U.S. DEPARTMENT OF COMMERCE
National Technical Information Service

PB-253 700

RESIDENTIAL ENVIRONMENTAL PREFERENCES AND CHOICE

SOME PRELIMINARY EMPIRICAL RESULTS RELEVANT TO URBAN FORM

REGIONAL SCIENCE RESEARCH INSTITUTE

PREPARED FOR
NATIONAL SCIENCE FOUNDATION

MARCH 1971

PB 253 700

RESIDENTIAL ENVIRONMENTAL PREFERENCES AND CHOICE: SOME PRELIMINARY EMPIRICAL RESULTS RELEVANT TO URBAN FORM

by

Mark D. Menchik

RSRI Discussion Paper Series:
No. 46

March 1971

The purpose of this series is to encourage the earliest possible dissemination of research results. Appearance of a paper in this series does not preclude later formal publication in journals or elsewhere. The author invites comments and reactions.

This investigation was supported by FWQA Grant No. 16090 DYX to the Regional Science Research Institute, and NSF Dissertation Support Grant GS-2817. The author is indebted for their helpful comments to Thomas R. Hammer, James C. Knox, and Robert E. Coughlin.

Regional Science Research Institute
GPO Box 8776
Philadelphia, Pennsylvania 19101

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA 22161

CONTENTS

		Page
0.	INTRODUCTION	1
	0.1. Discerning Preferences for a New Residential Environment	1
	0.2. Measuring Residential Preferences and Choice	4
	0.3. Preferences, Market Choice, and Demand-Oriented Theories of Urban Form	7
1.	THE QUESTIONNAIRE SURVEY	9
2.	DEFINITION AND PROPERTIES OF RESIDENTIAL PREFERENCE MEASURES	13
	2.1. Obtaining Residential Preference Data	13
	2.2. Properties of Residential Preference Measures	16
	2.3. Discussion of Residential Preference Measures	27
3.	SOCIO-ECONOMIC AND DEMOGRAPHIC INFLUENCES ON PREFERENCES	31
	3.1. Sex of Respondent	31
	3.2. Age of Household Head	34
	3.3. Family Income	36
	3.4. Childhood Residence	38
	3.5. Discussion: Interpreting Negative Results	40
4.	RESPONDENTS' RESIDENTIAL CHOICE	42
5.	RELATIONSHIPS AMONG RESIDENTIAL PREFERENCES AND CHOICE	52
	5.1. Definitions of, Relationships Among, the Preference Variables Used	52
	5.2. Relationships Among the Choice Variables	55
	5.3. Some Theory About First- and Second-Order Preference-Choice Relationships	57
	5.4. First-Order Preference-Choice Relationships	63
	5.5. Second-Order Preference-Choice Relationships	74
6.	SUMMARY OF RESULTS, CONCLUSIONS, AND SUGGESTIONS FOR FURTHER RESEARCH	80
	6.1. Results with Preference and Choice Variables	80
	6.2. Socio-Economic and Demographic Influences on Preferences	84
	6.3. Demand-Oriented Theories of Urban Form	85
	6.4. Evaluating a Cluster Development Pattern	86
	6.5. Suggestions for Further Research	88

INDEX OF TABLES AND MAP

		Page
Map 1.	Areas of Questionnaire	10
Table 1.	Number of Survey Respondents and Response Rate by Area	12
Table 2.	Considerations in Choice of a Place to Live	15
Table 3.	Frequency Distribution of Scores for Two Response Types	17
Table 4.	Residential Preferences, for Respondents Asked About Choice of Present or a Hypothetical Home	19
Table 5.	Interareal Comparison of Residential Preferences	21
Table 6.	Correlations Among Residential Preferences	25
Table 7.	Sex of Respondent and Residential Preference	32
Table 8.	Age of Household Head and Residential Preferences	35
Table 9.	Total Family Income (last year) and Residential Preferences	37
Table 10.	Childhood Residence of Respondent and Residential Preferences	39
Table 11.	Means and Standard Deviations for Residential Choice Variables	43
Table 12.	Product Moment Correlations Among Residential Environmental Variables	50
Table 13.	Residential Preference Variables	54
Table 14.	Axioms Generating Qualitative Correlation Matrix Between Preference and Choice Variables	65
Table 15.	Predicted Qualitative Preference-Choice Correlation Matrix	66
Table 16.	Product Moment Correlations of Residential Preference and Choice Variables	68
Table 17.	Predicted and Actual Second-Order Preference-Choice Correlations	77

ABSTRACT

In order to evaluate people's demands for new forms of residential environments (such as cluster development rather than the spread pattern of suburban sprawl), it may be useful to deal with persons' residential preferences directly, rather than their market choice. The paper develops residential environmental preference variables from questionnaire survey data. The preferences may be thought of as relative tradeoffs among the residence's accessibility, characteristics of the house and lot (e.g., lot size), the quality of the natural environment, and characteristics of the non-natural environment (e.g., population density). At the same time, comparable measures are devised of the characteristics of the persons' present residences --- their residential choice.

Principal results are that the preference and choice variables do seem useful; that preferences, thus defined, do express themselves to some extent through market choice; and that different persons do in fact prefer different residential characteristics. That is, there are persons with high natural environmental preferences who are willing to sacrifice a large lot to live in a beautiful area.

RESIDENTIAL ENVIRONMENTAL PREFERENCES AND CHOICE: SOME PRELIMINARY EMPIRICAL RESULTS RELEVANT TO URBAN FORM

0. INTRODUCTION

The research presented below is an attempt to devise quantitative measures of persons' preferences for different aspects of their residences--including the quality of the natural environment--as well as quantitative measures of the characteristics of their present residences. Persons' preferences are related to their residential choices. A fundamental motivation for this work is that the measures devised should ultimately be relevant to some of today's pressing questions of the 'most desirable' or 'least environmentally damaging' forms of metropolitan growth. 'Preferences' and 'natural environmental quality' are both often vague and general concepts; explanation of them and the purposes of the present effort should justify a lengthy introduction.

0.1. Discerning Preferences for a New Residential Environment

Some of the data and ideas below had their origin in the Upper East Branch Brandywine Project (see [9]), an attempt to apply the new legal mechanism of 'conservation easements' to protect a stream basin from the environmental and water quality degradation accompanying conventional suburbanization. In place of the usual uniformly low population density spread or 'sprawl' pattern of suburban development, the plan would

accommodate projected growth in a 'cluster development' pattern. Residential construction would be restricted in environmentally (particularly hydrologically) sensitive or important areas such as flood plains, steep slopes or forests, and be concentrated elsewhere. An immediate and important question is whether persons would want to live in the residential environment envisioned by the Brandywine Plan, rather than conventional suburban or urban areas.

Determining the desirability of a new residential environment is clearly a difficult task, parallel to the job of calculating the demand for a new commodity. Quandt and Baumol [20] estimated the demand for a new transport mode by considering any possible form of travel, or an 'abstract transport mode', as a combination of the important attributes of travel time, trip cost, and departure frequency. Similarly, we may define a residence as a collection of attributes, or a residential bundle (Herbert and Stevens [8]). For present purposes, important elements of the residential bundle include accessibilities, characteristics of the house and lot, the quality of the natural environment, and characteristics of the nonnatural or man-made environment such as local population density, the neighborhood's social composition and the provision of facilities like schools.

Now a residence has many attributes, and there are many ways of looking at an individual's residential choice, but it is suggested that the above categorization is helpful for present purposes. This is because these categories not only represent

characteristics assumed important in individuals' residential location decision, but because this categorization is deemed useful in the large scale comparison of different kinds of residential bundles, present and future. Thus, moving outwards from the center of a large metropolitan area we find that, on the average, subareas have reduced generalized accessibility, larger houses and lots, increasing quality of the natural environment, and systematic changes in the man-made environment, such as reduced population densities and different kinds of neighbors. Similarly, compare the Brandywine cluster plan with the usual outlying suburban development. Both would have similar generalized accessibility, but the former would have smaller house lots, a higher quality natural environment, and higher nearby population density, since a smaller fraction of land is in residential land use. Accordingly, a knowledge of persons' relative preferences for lot size, natural environmental quality, and population density might enable us to predict the demand for either of the two residential environments.

Thus, defining persons' preferences according to the above categorization may be one way to begin assessing preferences for new residential environments.[1]

---

[1]Of course, preferences are learned, and while they influence individuals' market behavior they are in turn influenced by market offerings, persons' housing experiences. As is common in preference analysis, we will be assuming away any influence that the opportunities for choice, and the making of the choice itself, has on preferences.

0.2. Measuring Residential Preferences and Choice

A second, related motivation for this research is to cast some light on the measurement of residential preferences. 'Preference' will initially be defined to indicate some relative weighting of the desirability of alternative residential attributes.

One way to measure residential preferences is by their expression in housing market choice. A difficulty of this approach is indicated by the frequent opinion that the expression of one's preferences in the housing market is quite limited. The limitation does not come from the fact that most persons cannot afford to live in their 'dream home', for this is true of any economic commodity, and preferences are being defined in terms of relative desirabilities. Rather, some feel that housing opportunities are less diverse than people would like, especially in terms of combinations of items in the residential bundle, rather than just house design characteristics. Thus, as has been mentioned when we look at areal averages, we see that accessibility, house and lot characteristics, natural environmental quality, and characteristics of the nonnatural environment all vary systematically as one moves outwards from the urban core. A person desiring a non-standard combination of these residential attributes, such as that proposed in the Brandywine plan, would have difficulty expressing his preferences in the market.[2]

---

[2]It should be emphasized that the difficulty of inferring preferences from market choice is partly a data problem. Given data on small areal units or, ideally, individual residences, the variety of combinations of residential attributes should increase markedly. For example, a small home or apartment in area that is predominated by large lots would satisfy a preference for low density surrounding, but a slight preference for a large lot.

Direct measurement of residential preferences has often proceeded by having persons in an experimental context choose among several residential alternatives. Alternatively persons may state a quantitative degree of preference for a residence, a residential environment, or an attribute, as being close to work (see, e.g., Peterson [19]). Or, within a questionnaire survey one may ask persons structured or unstructured questions about residential preferences, or their actual choice of a place to live. The data used below were derived from an unstructured question asked during a questionnaire survey and analyzed according to the categories discussed. A fundamental difference, accordingly, between the present effort and much of the other residential preference work is that here most of the important variables and categories have been defined on an a priori, nonstatistical basis, as being useful for urban form questions. Many other preference studies arrive at preference variables and categories by multivariate statistical procedures (e.g., principal component or factor analysis, or grouping techniques), in a posteriori fashion.

Without attempting a detailed evaluation of alternative means of collecting preference data, one may state that an advantage of a survey approach is that it would tend to produce samples of persons that are more 'random', 'representative' or, certainly more 'controlled' in their residential, socio-economic, and demographic characteristics than appears to be the case when the subject must be brought into an experimental situation.

Furthermore, survey techniques would seem to be more suited to obtaining large samples than experimental procedures. Indeed, a goal of the present research is to see if an unstructured survey question yields useful residential preference information.

Since this research emphasizes the relative importance of natural environmental and other residential preferences, a related goal is whether the vague concept of 'natural environmental preferences' may be defined sufficiently well enough to obtain useful results. At the same time, we will see how the 'natural environmental quality' of the survey respondents' residences may be defined and measured.

Notice the present research concentrates not on natural environmental perception but preferences. The attention paid to the former topic is often justified by the reasoning that since we know so little about man's response to his physical environment, the first step should be to map out the dimensions of environmental perception without initially measuring persons' preferences on these dimensions. A different approach is taken here: perceptual dimensions and variables are not dealt with explicitly. Instead we see if our measures of the natural environmental quality of persons' residential environments are linked sufficiently well to their residential preferences to suggest that the 'quality' measures are in turn related to how people perceive their residential environments.

0.3. Preferences, Market Choice, and Demand-Oriented Theories of Urban Form

The third major focus of this research is to see whether our preference measures are related to the measures of the characteristics of persons' residences, that is to see whether persons' preferences are associated with their market choices. Clearly, there should be some empirical relationship between preference and choice. If not, doubt is cast on the theoretical usefulness of a preference approach to residential choice, or (much more likely) on the validity of our preference and choice measures.[3]

The theoretical and empirical research work on urban form of Alonso [1], Herbert and Stevens [8], Harris et. al. [6], and Muth [17] (among others) is similar in its orientation to residential preferences and thus the demand side of the urban housing and land markets, frequently to the exclusion of the market supply characteristics. This demand orientation is justified by the assumption that, in the long- or short-run, supply will respond to demand. Assessment of the existence of a preference-choice relationship thus may cast some slight empirical light on the reasonableness of such assumptions.

It is important not to over-emphasize this point. Testing for a preference-choice relationship can in no way be said to 'test' these general and abstract theories: the operation is largely a test of the validity of the preference and choice

---

[3]Just as clearly, though, we should not expect a strong preference-choice relationship. See the discussion in subsection 5.3 below.

measures adopted. Thus, if no preference-choice relation were apparent in the data, the theories could be perfectly correct, but the measures chosen could be quite poor. More interesting is the possibility that only a very weak preference-choice relationship be found, but one that is definitely statistically significant.[4] Then one might infer that yes, residential preferences do influence residential choice, but the weakness of the relationship suggests that supply considerations, information as to residential opportunities, and other influences that were assumed away in the theory should in fact be included. Such an interpretation is quite dangerous, for it is more reasonable to assume that weak relationships are caused by imperfections in the measures used.

A final point: this research is quite preliminary. The variables used could hardly be considered 'refined' for, indeed, an objective of this work is to see whether results using unrefined data are of sufficient interest to encourage undertaking the effort of refining the variables. Similarly, the statistical analysis is in many cases incomplete. Other variables could have been used in the analysis and some statistical procedures were selected largely because of their simplicity. Hopefully, these preliminary results should stimulate further work in this area that is as important as it is difficult.

---

[4] This is, in fact, our finding.

Below, Section 1 describes the sampling and interviewing procedures used, while Section 2 discusses the definition of the residential preference variables and their statistical properties, justifying the conclusion that the data obtained appear to be useful preference measures. In Section 3 respondents' residential preferences are related to several socio-economic and demographic variables, with generally negative results. The variables describing the characteristics of respondents' residences are defined and discussed in Section 4. The existence of a relationship between preference and choice is found in Section 5. Two kinds of preference-choice relationships are defined there, the second one being sensitive to the (inherently arbitrary) scale units employed. However, this difficulty may be overcome. A summary of results, conclusions about the variables used, the nature of residential preferences, theoretical questions raised, and implications for urban form policy appear in Section 6, as well as suggestions for further research.

1. THE QUESTIONNAIRE SURVEY

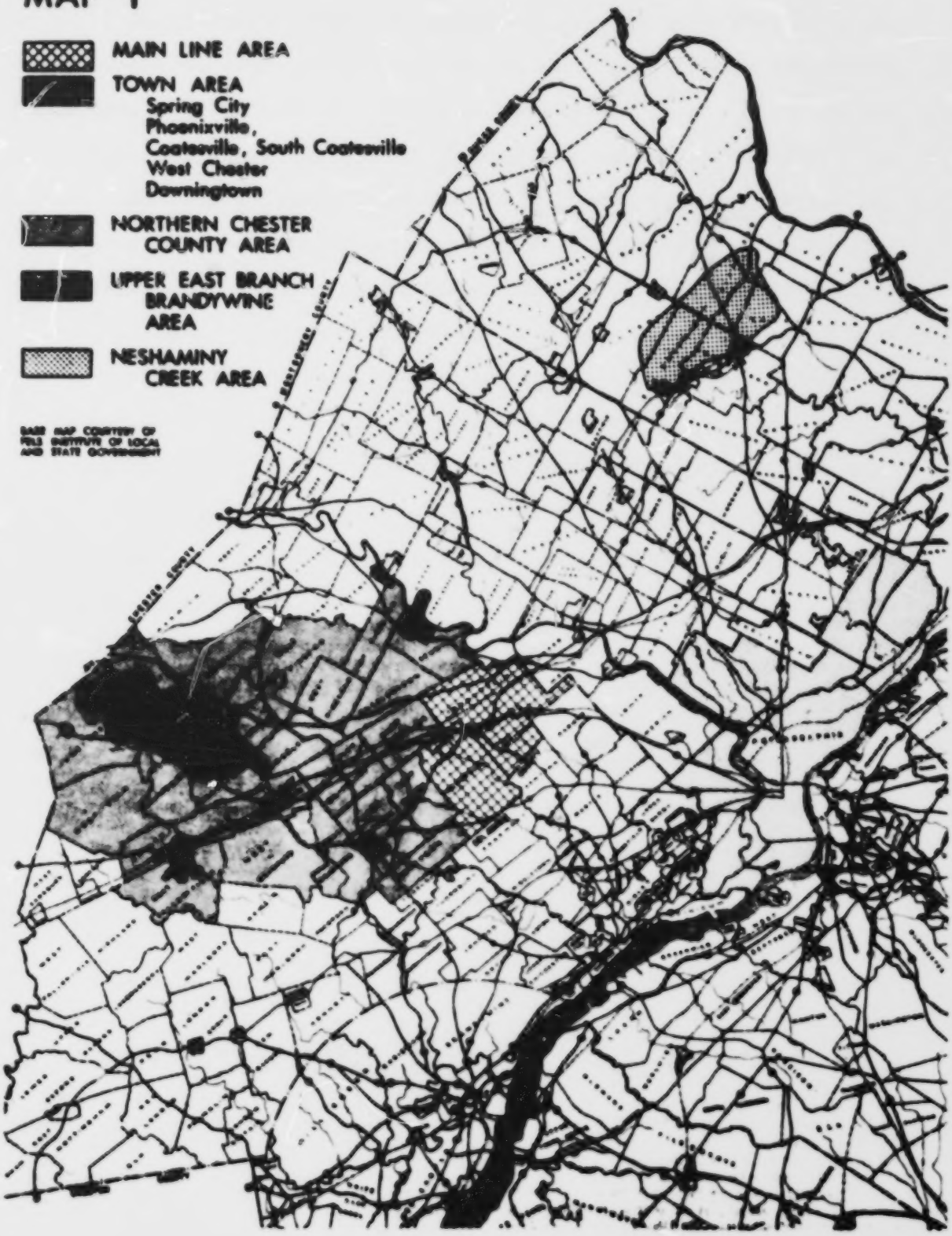
A home interview questionnaire survey with 457 respondents was conducted in October and November 1966.[5] Using a cluster sampling design (see, e.g., Keeping [10, pp. 155-157]), samples were drawn at different sampling rates in each of five areas in southeastern Pennsylvania (see map 1): (1) the Upper East Branch Brandywine Creek area, a predominantly rural area in

---

[5]See Menchik [15] for a description of other aspects of the survey, and a copy of the questionnaire form.

AREAS OF QUESTIONNAIRE

MAP 1



western Chester County; (2) the Neshaminy Creek area, a predominantly rural area in upper Bucks County; (3) the Northern Chester County area, rural and suburban; (4) the Chester County Towns area, several relatively densely settled towns; and (5) the Main Line area of northeastern Chester County, a relatively wealthy and recently subdivided suburban area. Both Bucks and Chester are suburban counties within the Philadelphia Standard Metropolitan Statistical Area (SMSA).

Table 1 indicates (second row) the number of respondents residing in each area. The survey response rate, the proportion of attempts to interview that resulted in completed interviews, is 71% overall, sufficient for some confidence in the representativeness of responses, and is similar in all the areas surveyed ($\chi^2 = 4.813$, $df = 4$, $.50 < p < .30$). The interviewing procedure was designed to avoid biassing the sample towards respondents of one sex: where the household contained both male and female heads (commonly, husband and wife), a random mechanism ensured that the male head would be interviewed with probability one-half.

In virtually all the analysis below, respondents from all areas are aggregated, and treated as one sample of 457. Since different sampling rates were used in each of the five areas, this unweighted 'sample' is in fact representative of no single geographic area. The aggregation of respondents was done, however, to allow a sample size adequate for statistical analysis. Possible inferential difficulties caused by this aggregation are evaluated in Section 2.

Table 1. Number of Survey Respondents and Response Rate, by Area

	Upper East Branch Brandywine		Neshaminy Creek		Northern Chester Co.		Towns		Main Line		Total	
	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
Occupied Dwelling Units Visited	225	100%	150	100%	130	100%	72	100%	68	100%	645	100%
Completed Interviews	165	73	109	73	89	68	44	61	50	74	457	71
Respondent Repeatedly Absent	39	17	11	7	25	19	21	29	16	24	112	17
Respondent Refused	17	8	23	15	14	11	4	6	1	1	59	9
Other	4	2	7	5	2	2	3	4	1	1	17	3

2. DEFINITION AND PROPERTIES OF RESIDENTIAL PREFERENCE MEASURES

2.1. Obtaining Residential Preference Data

Respondents were asked one of two series of questions about the choice of a place to live. When the respondent had lived at his present home for 10 years or less and that home had been chosen by the respondent or spouse, then the respondent was asked about the choice of the present home. The question was phrased: "When deciding on a new home in a new area, some people judge a house and area by one set of characteristics, while others judge them by other characteristics. When you were trying to decide on a place to live what were all of the characteristics that were important to you?" The respondent's open-ended reply was written down verbatim by the interviewer.

The remaining respondents, those whose present residence had not been chosen by them, or had lived there more than ten years, were asked about a hypothetical move: "When deciding on a new home in a new area, some people judge a house and area by one set of characteristics, while others judge them by other characteristics. Suppose you were trying to decide on a place to live. What would be all of the characteristics that would be important to you?" Again, the response was copied word-for-word.

The residential preference data used below were obtained from a content analysis of the replies to these questions. The researcher looked through some of the questionnaires, and established 26 response types, based on relevance to the replies

and usefulness for analysis. The response types are listed in Table 2. The coding procedure and definitions of the response types were detailed in a coding manual that was used to train two coders. The coders read through the respondent's reply, counting the number of times the respondent made reference to each response type and the residential locational consideration that it represents. To check inter-coder reliability,[6] the two coders coded the same set of 100 questionnaires. Inspection shows substantial agreement between the coders, especially with respect to the four major response categories of Table 2.

The respondent's preferences for any item (or combination of items) in the residential bundle is operationally defined by the number of times the relevant response type(s) was mentioned. The questions were designed to elicit considerations important in the choice of a home (present or hypothetical) and, given the questions' unstructured form, it was assumed that the content of the response would indicate the degree of importance or preference for any locational consideration. Furthermore, the relative response frequencies of two response types was assumed to indicate relative preferences; a respondent mentioning the design of the home and not the natural environment is

---

[6]The correlation between the scores obtained by two coders on the same set of questionnaires.

Table 2. Considerations in Choice of a Place to Live (Number and Percentage of Respondents Making One or More Responses of Each Type)

	No. of Respondents	% of Respondents
Natural Environmental Beauty (B)		
1. Ruralness, woodsiness, country-like character of the area	103	22.9%
2. Specific mention of an area's prettiness, beauty of surroundings, view	23	5.1
3. Other specific considerations referring to natural environment: climate, weather, topography, etc.	46	10.2
Man-Made or Non-Natural Environment (M)		
Density: presence and number of persons in the vicinity, in general, without reference to any of their characteristics		
4. Good: Wanted to be near people	6	1.3
5. Bad: Didn't want to be near (too many) other people	44	9.8
6. Neutral: Just interested in whether there are many persons nearby	2	0.4
Characteristics of people nearby		
7. Good: Wanted to be near a particular kind of person, near my kind of people, etc.	78	17.4
8. Bad: Don't want to live near some stated kind of person, disrespectable, unfriendly, etc.	16	3.6
9. Neutral: Wanted to know what kinds of persons around	3	0.7
Presence of traffic and non-residential land uses like stores, factories, etc.		
10. Good: Wanted to live near a store, etc.	4	0.9
11. Bad: Didn't want stores, factories, etc., wanted purely residential, little traffic	21	4.7

Table 2. (Continued)

	No. of Respondents	% of Respondents
12. Neutral: Interest in presence of stores in vicinity	2	0.4
13. Other considerations referring to the non-natural environment: nice neighborhood, good section of town, quiet, noise, privacy, well-paved streets, etc.	87	19.4
Accessibility (A)		
14. To work	63	14.0
15. To schools	73	16.3
16. To churches	46	10.2
17. To transportation facilities	45	10.0
18. To shopping	69	15.4
19. To other specified activity or convenience, in general	70	15.6
Characteristics of House and Lot (H)		
20. House size, number of rooms, size of rooms	112	24.9
21. Price of house and land, purchasing or renting arrangements, mortgage, etc.	54	12.0
22. House design characteristics, specific and general, room layout, heating, plumbing, etc.	189	42.1
23. Lot size	94	20.9
24. Other natural environment lot characteristics: stream in backyard, soil, landscaping, etc. lot itself, not vicinity	39	8.7
25. Other non-natural environment lot characteristics: distance from road, sewage facilities, orientation, etc.	30	6.7
Other		
26. Respondent or spouse used to area, lived here before	12	2.7
No. of respondents with usable answers	449	
Respondents with no usable answers	8	

assumed to prefer house design considerations more than those of the natural environment. Given an unstructured response situation (i.e., like those of the Thematic Apperception or Rorschach projective tests, see Colby and Menchik [26]) it is common in personality and social psychology to define the mention of some item as indicating the individual's orientation toward that item. See, for example, McClelland's definition and measurement of the need-for-achievement motive [14].

2.2. Properties of Residential Preference Measures

The remainder of this section discusses some of the properties of the preference measures just defined. While the raw preference data for each individual consists of the number of times each response type is mentioned, only rarely is a given response type mentioned more than once. Table 3 presents the frequency distribution for two response types used later: the infrequently mentioned p_c^*, dislike of densely populated surroundings (response type 5 in Table 2), and the frequently mentioned p_l^*, house lot size considerations (response type 23 in Table 2). In both cases the variables are effectively binary.

In Section 3, and the remainder of this section, the preference data will be analyzed in terms of the four major categories of Table 2, each a combination of response types. The major preference categories are accessibility considerations (A), characteristics of the house and lot (H), natural environmental beauty (B), and characteristics of the nonnatural or man-made environment (M). As mentioned in the introduction, these

Table 3. Frequency Distribution of Scores for Two Response Types

No. of mentions of that response type	P_c^* 5. Dislike of Dense Surroundings (No. of Respondents)	P_ℓ^* 23. Lot Size (No. of Respondents)
0	405	355
1	42	90
2	2	3
3	0	1
Total No. of Respondents	449	449

categories were thought relevant to urban form considerations. Hence, they are defined on a prior or nonstatistical basis.[7]

The last category (M) mixes two aspects of the nonnatural environment that for certain purposes should probably have been kept distinct: (1) indications of dislike of high population density in the surroundings (response type 5) and the noise, traffic, etc. associated with population density (part of response type 13) versus (2) considerations of land uses in the vicinity and the kinds (not numbers) of people there (the remaining response types plus part of response type 13).

An immediate question to be asked of the preference measures is whether the responses appear to be influenced by the form of the questions used (see above), since some respondents were asked about the choice of the present home, others about a hypothetical move. Table 4 shows that the frequency of responses in the four major categories, or combinations of response categories does not vary with the form of the question asked. For no response category or combination was the difference in frequencies for the two forms of the question statistically significant at even the $p = .10$ level. Examination of responses for individual areas also indicates similar response frequencies for the two forms of the question. For this reason, responses for the two question forms have been pooled for all further analysis.

---

[7] Data from respondents in the Towns area are omitted for the tables of Section 3 and the remainder of Section 2.

Table 4. Residential Preferences, for Respondents Asked About Choice of Present or a Hypothetical Home

(Number and Percentage of Respondents Making One or More Responses for each Major Category, or Combination of Categories.)

Category	Present Home		Hypothetical Home		Total	
	No.	%	No.	%	No.	%
A (Accessibility)	125	49%	78	52%	203	50%
H (House and lot)	172	67	104	69	276	68
B (Natural Environment)	88	34	56	37	144	35
M (Non-Natural Environment)	119	46	76	50	195	48
AH(Access.-House)	67	26	48	32	115	28
AB(Access.-Nat. Env.)	44	17	27	18	71	17
AM(Access.-Non-Nat. Env.)	58	23	44	29	102	25
HB(House-Nat. Env.)	48	19	30	20	78	19
HM(House-Non-Nat. Env.)	62	24	42	28	104	25
BM(Nat. Env.-Non-Nat. Env.)	52	20	30	20	82	20
Total No. of Respondents	257	100	151	100	408	100
No Usable Answers	2		3		5	

Whether inter-areal preference differences exist is important because, as mentioned, the areas were sampled at different rates and all respondents are pooled together for most of the analysis. If preferences differed importantly among areas, then it might be hazardous to draw conclusions for the entire sample based on preference frequencies, because such inferences might be due solely to the weight each area was given in the total sample. For example, if all rural people strongly preferred the natural environment, and no suburban or central city residents did, then a strong preference for the natural environment found in the total sample might be simply due to the overrepresentation of rural areas in the total sample.

Table 5a shows, by area, the number of respondents making one or more response for any major preference category. The base used for calculating percentages is the total number of respondents in each area giving usable preference responses. The frequencies in each row of Table 5a form a 2 X 4 contingency table (e.g., stating A or not stating A, versus the 4 areas) for which the χ^2 statistic appears. For preference category B the percentage of respondents ranges from 20% to 45%, and these interareal differences are highly significant statistically.

However, notice interareal differences in the <u>overall</u> response frequencies. Summing down the first column of Table 5b observe that in the Brandywine there were a total of 356 major preference categories 'responded to', that is, major preference categories having one or more responses: Brandywine respondents thus 'responded to' an average of 2.18 response

Table 5. Inter Areal Comparison of Residential Preferences

a. Number and Percentage of Respondents Making One or More Responses for each Major Category, or Combination of Categories

Preference Category	χ^2	Brandywine No.	Brandywine %	Neshaminy No.	Neshaminy %	Northern Chester Co. No.	Northern Chester Co. %	Main Line No.	Main Line %	Total No.	Total %	Approximate 95% Confidence Intervals	
A(Accessibility)	6.368+	91	56%	45	42%	46	52%	21	42%	203	50%	45%	55%
H(House and lot)	5.273	103	63	72	67	68	78	33	66	276	68	62.5%	72.5%
B(Nat. Env.)	12.950**	73	45	34	32	27	31	10	20	144	35	31%	40%
M(Non-Nat. Env.)	6.440+	89	55	42	39	42	48	22	44	195	48	43%	53%
AH(Access.-House)		49	30	25	23	31	35	10	20	115	28		
AB(Access.-Nat. Env.)		39	24	13	12	14	16	5	10	71	17		
AM(Access.-Non-Nat. Env.)		55	34	17	16	20	23	10	20	102	25		
HB(House-Nat. Env.)		41	25	15	14	16	18	6	12	78	19		
HM(House-Non-Nat. Env.)		44	27	19	18	30	34	11	22	104	25		
BM(Nat.Env.-Non-Nat.Env.)		38	23	19	18	18	21	7	14	82	20		
Total No. of Respondents		163	100	107	100	88	100	50	100	408	100		
No Usable Answers		2		2		1		0		5			

df = 3, +interareal differences, statistically significant $p < .10$
**interareal differences, statistically significant $p < .01$

Table 5. (Continued)

b. Number of Respondents Making One or More Responses for each Major Category and Percentage of Total Number of Major Categories Responded to (See Text).

Preference Category	Brandywine No.	%	Neshaminy No.	%	Northern Chester Co. No.	%	Main Line No.	%	Total No.	%	Approximate 95% Confidence Intervals	
A(Accessibility)	91	26%	45	23%	46	25%	21	24%	203	25%	21%	28%
H(House and lot)	103	29	72	37	68	37	33	38	276	34	30%	38%
B(Natural Environ.)	73	21	34	18	27	15	10	12	144	18	15%	21%
M(Non-Nat. Env.)	89	25	42	22	42	23	22	26	195	24	20%	27.5%
Total	356	100	193	100	183	100	86	100	818	100		
No. of Respondents	163		107		88		50		408			
Average no. of categories responded to per respondent												

$df = 9, \quad \chi^2 = 9.626, \quad .30 < p < .50$

categories, while for the Main Line this figure is 1.72. This suggests a reason for the percentages in the first four rows of Table 5a being generally lower for the Main Line than the Brandywine: the former respondents make fewer responses, in general, than the latter. Accordingly, percentages in Table 5b are computed from the base of the total number of major categories 'responded to'. Notice that this reanalysis markedly reduces the strength of the interareal differences noted in Table 5a: for category B the range becomes 12% to 21%. The frequencies in Table 5b form a 4 X 4 contingency table (4 major categories versus areas) where the interareal preference differences are not statistically significant, as shown at the bottom of the table. A chi-squared test may be performed for just the B category (B responses and non-B responses versus areas) where the marginal totals are major categories responded to, not numbers of respondents, as in the chi-squared tests of Table 5a. For this test, df=3, $\chi^2=5.208$, $.10 < p < .20$.

Thus, because the average number of major categories responded to varies interareally, results of an interareal preference comparison depend on whether the base used is 'respondents' or 'categories responded to'. With the former base, there are marked differences in natural environmental preferences, which are statistically significant. For the latter base, the magnitude of the differences decrease, and statistical significance vanishes. Either base may be appropriate, depending on the purposes of analysis.

A further question pertains to interareal differences in preference relationships, either among the preference

categories, or between preference and other variables. While the frequencies of B and M preferences might vary across areas, the relationship between the two variables might not, so that we could use the total sample to infer the relationships among preferences for some larger population.

The data of Table 5 may be used to compute 2 X 2 contingency tables for each area, and each pair of major categories. For example, one cell of the AH contingency table would indicate the number of respondents responding to both the A and H categories. The chi-squared statistic computed from such tables tests the statistical significance of the relationship between each pair of variables, that is whether mention of A significantly raises (or lowers) the probability of mentioning H. It is furtner possible to compute a measure of the strength of the association, or degree of correlation, between each pair of binary variables. This is the contingency coefficient, C, where

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

and N is the total number of observations in the table (see Siegel [23, pp. 196-202]). The contingency coefficient C is quite different from other measures like the product-moment or rank-order correlation coefficients, and is not directly comparable to them. C may be computed from data scaled only nominally, appearing in the form of contingency tables, and is thus a 'less powerful' measure of correlation than the other statistics. For 2 X 2 tables, C has an upper limit of 0.707 ... . While C is ordinarily nonnegative, the use of

signs for the 2 X 2 tables in Table 6 has the obvious interpretation: a negative C for AH indicates that mention of A reduces the probability of mentioning H. Further, in computing χ^2 for Table 6, no corrections for continuity were made (see Siegel [23, p. 107]). All in all, contingency coefficients represent convenient measures of association for these binary variables. Significance tests on C and χ^2 are equivalent.

Table 6a shows the C and associated χ^2 values for 6 combinations of major preference categories, 4 areas, and the total sample. The relationships among preference categories are shown to be generally 'the same' interareally, with just 5 exceptions (underlined in Table 6a) out of 24 cases. By the relationships being 'the same' it is meant: (1) if C in the total sample is non-significant, none of the C's for individual areas should be statistically significant (no exceptions) and (2) if C in the total sample is statistically significant, those for individual areas should also be significant, and with the same sign. In the latter case of the 5 exceptions, 4 had the correct sign but failed significance, while only one (Brandywine, BM) had the incorrect sign. Notice also that the same general level of C obtains across the rows of Table 6a; this is especially true for AH and (except for Brandywine) BM.

Thus, relationships among the major preference categories appear quite stable interareally, suggesting that such relationships are a fit object for further study.

The relationships of Table 6, though, represent much aggregation from the 26 original response types of Table 2.

Table 6. Correlations Among Residential Preference Categories

a. Interareal Comparison of Preference Correlations (underlined cells are exceptions to that row's relationship; see text)

Preference Categories	Brandywine C	Brandywine χ^2	Wesh&miny C	Wesh&miny χ^2	N. Chester Co. C	N. Chester Co. χ^2	Main Line C	Main Line χ^2	Total C	Total χ^2
AH(Access.-House)	-.213	7.733**	-.206	4.858*	-.240	5.959*	-.314	5.451*	-.228	22.324
AB(Access.-Nat. Env.)	-.042	0.310	-.053	0.298	-.006	0.003	-.081	0.328	-.000	0.018
AN(Access.-Non-Nat. Env.)	+.130	2.633	-.026	0.071	-.089	0.697	+.062	0.192	+.048	0.974
HB(House-Nat. Env.)	-.130	2.606	-.319	12.157**	-.275	7.197**	-.062	0.200	-.208	18.478**
HN(House-Non-Nat. Env.)	-.298	15.334**	-.353	15.274**	-.132	1.562	-.287	4.482*	-.281	34.967**
BN(Nat. Env.-Non-Nat. Env.)	-.046	0.346	+.226	5.780*	+.245	5.600**	+.253	3.429	+.134	7.468**
No. of observations	163		107		88		50		408	

df = 1 *statistically significant, $p < .05$
**statistically significant, $p < .01$

(b) Signs of Preference Correlations for Total Sample

	A	H	B	N
A		-	0	0
H			-	-
B				+
N				

A product-moment correlation matrix (not shown) was computed for the 26 response types, using the number of times that response type was mentioned, and including respondents from the Towns area. It was difficult to make reliable inferences from that correlation matrix because of the very low frequency of occurrence of some response types (see Table 2), and the possibly related fact that the correlation coefficients were quite small, a problem that will appear again below. Except for correlations of response types within major categories, only rarely was the absolute value of r greater than .20.

Even so, inspection of that correlation matrix suggests relationships of the same signs as in Table 6b, this time, of course, among the response types contained in each major response category.

Since the major categories were defined from the response types on a prior, nonstatistical basis, it is interesting to see how intercorrelated are the response types within each major category. Uniformly high and positive correlations among the responses are intuitively good indications of the 'homogeneity' of the major categories. Furthermore, in such a situation it is well known that the first principal component would explain a large fraction of the variance and would correspond to an unweighted summing of the original variables, just as was used in constructing the major categories (Morrison [16, p. 244])..

Again, inferences are difficult but it appears that the response types in categories A and H are positively intercorrelated, and strongly related compared to the other coefficients in the correlation matrix. The response types in

B and M are less strongly, but still positively associated.

2.3. Discussion of Residential Preference Measures

Having considered some of the statistical properties of the preference measures, let us draw some substantive conclusions about residential preferences from the preceding analysis. What are the relative preferences for the 4 major preference categories: accessibility (A), characteristics of house and lot (H), characteristics of the natural environment (B), and characteristics of the non-natural environment (M)? To measure 'strength of preferences' one may use the percentages in Tables 5a and 5b and observe that the ordering of percentages is the same for all areas. Characteristics of house and lot are mentioned by the most respondents, followed by accessibility and characteristics of the nonnatural environment, mentioned by almost identical numbers of respondents. Least preferred are characteristics of the natural environment. The ordering of H, A & M, B is statistically significant in the sense that the approximate 95% confidence intervals of Table 5a (computed using the tables in [18, p. 228]) are nonoverlapping, and the confidence intervals in Table 5b overlap minimally.

The primacy accorded to house and lot characteristics is consistent with the voluminous survey literature (generally derived from structured questions) on residential locational preferences and the decision to move. (See, e.g., Gans [5, p. 35], Lansing and Mueller [12, p. 21], and Rossi [23, p. 154]).

There are four reasons, though, why it is thought that this ranking of residential characteristics should be inter-

preted very carefully. First, even though the ranking is uniform throughout the four areas studied, the total sample is clearly not representative of a whole metropolitan area, nor are individual areas even representative of, say, 'the suburbs'. As mentioned, inferences pertaining to individuals or relationships <u>among</u> preferences are more likely to be generalizable to whole metropolitan areas, than descriptive summaries.

Second, we are not dealing with preferences for standardized, homogeneous commodities; the preferences for the residential categories used here are intimately dependent on the definition of those categories, which is ultimately arbitrary. For example, given more major categories than four, the ordering among the original four might change. As a given preference category is made 'wider' or more inclusive, one would expect it to be mentioned more often. The ill-definition of preference categories in the present situation is thus distinguished from the theory of consumer behavior, or decision theory in general, where preference functions are defined on a well-defined set of alternatives or a choice space. The situations are similar, though, inasmuch as in decision theory preferences defined over different sets of alternatives are not directly comparable.

Third is the problem of units and 'quantities' of the residential characteristics. Even if we were working with standardized residential characteristics like the acreage of homogeneous land, we would have no idea what lot size the respondent had in mind when he expressed a preference for a 'large' lot. Perhaps the low preference for the natural

environment occurs because the respondents do not have experience with thoroughly beautiful natural environments. If they had such an environment in mind, might they not express more preference for the natural environment? Again, the theory of consumer behavior, and decision theory in general, avoids this problem by working with well-defined quantities of the items concerned. There, preferences are defined relative to the units involved and changing units can alter the preference orderings.

The fourth point is that the presentation of the data in Table 5 is not a statistically 'efficient' means of summarizing individuals' rankings, for it does not work with individual rankings, but response frequencies. Using the number of responses for each major category, we could rank the four categories for each individual, allowing ties. There are many ways to aggregate such rankings, and a vast literature on this problem (Luce and Raiffa [13, ch. 14]). Taking the 'average ranking' for each major category, for example, has been shown to have certain desirable properties (Kendall [11, p. 87]). The coefficient of concordance, W, would then indicate the degree of agreement among the individuals (Siegel [23, pp. 229-238]).

All this, however, is not to discourage any simple inferences about strengths of preferences; it is to indicate the assumptions and definitions on which such inferences are covertly based, and pinpoint the difficulties of comparison involved. The problem of the unit of measurement will recur below (see subsection 5.3), where some ways of dealing with it will be discussed.

Another conclusion to be derived from the data pertains to interareal preference differences. Table 5 shows that such differences may not exist, or be substantial, for the preferences in general, though they may be important for natural environmental preferences. However, it is useful to ask why we are interested in interareal preference comparisons. The areas differ in socio-economic and demographic characteristics that, presumably, are responsible for preference differences, and also differ in the residential environments that are likely to result from preference differences. Given data for these variables on individual respondents, it is better to analyze the individual data directly, rather than compare whole areas and run the risk of 'ecological correlations' among areal averages that do not obtain for individuals (see, e.g., Duncan, Cuzzort, and Duncan [3]). Direct examination is done in section 3 below.

It is quite useful, though, to draw conclusions about preference relationships, using the results of Table 6. Except for the relationship between natural environmental and nonnatural environmental preferences, all the major preference categories are either negatively associated, or uncorrelated. It thus appears that the residential characteristics of accessibility, house and lot, and natural and nonnatural environments (the last two taken as one) are distinct entities. These characteristics may be thought of as similar to substitute commodities in the sense that if a person has a preference for one he is (frequently) less likely to prefer another, or at least not more likely.

It is difficult to draw conclusions about the positive

relationship between B and M. The latter category contains density considerations as well as concerns about kinds of persons, land uses, etc. It would seem that persons concerned about the quality of the natural environment might manifest this concern by a preference for the avoidance of densely settled or noisy areas, but there might be no necessarily positive association between B and the other items in M. Does the fact that the contingency coefficient between B and M is negative for the Brandywine indicate different proportions of the two components of M, or a different relationship between preferences for the natural environment and for low density? Perhaps Brandywine residents, through acquaintance with a pretty natural environment and low population density, have learned to distinguish the two? At present, we cannot tell.

3. SOCIO-ECONOMIC AND DEMOGRAPHIC INFLUENCES ON PREFERENCES

This section considers the relationship, for individual respondents, between the 4 major preference categories of the previous section and the sex of the respondent, age of household head, total family income, and type of place of the respondent's childhood residence. The latter set of variables (especially the first three) are among those considered 'socio-economic' and 'demographic' and are routinely used as possible explanations for residential and environmental preferences and choice (see, e.g., Sonnenfeld [24]).

3.1. Sex of Respondent

Table 7 shows frequency of mention of each of the four major preference categories, by sex of respondent. Percentages

Table 7. Sex of Respondent and Residential Preferences

Preference Categories	Male No.	Male % of Respondents	Male % of Categories Responded to	Female No.	Female % of Respondents	Female % of Categories Responded to	Total No.	Total % of Respondents	Total % of Categories Responded to
A(Accessibility)	71	43%	22%	132	54%	27%	203	50%	25%
H(House and lot)	110	67	34	166	68	34	276	68	34
B(Nat. Env.)	63	38	20	81	33	16	144	35	18
M(Non-Nat. Env.)	79	48	24	115	47	23	195	48	24
Total Respondents	164	100		244	100		408	100	
Total Categories Responded to	323		100	495		100	818		100

Significance test for entire contingency table (see text):

$df = 3,\ \chi^2 = 2.968,\ .30 < p < .50,\quad C = .060$

in that table are calculated, first, to the base of the total number of respondents for that column (that is, all males, all females, or the total number of respondents in the survey). Second, percentages are calculated using as a base the total number of categories responded to. These two forms of percentages are those used in Tables 5a and 5b.

No prior hypotheses were established for the variables in Table 7. The fact that accessibility is mentioned by a higher proportion of female than male respondents, though, makes sense because of the wide variety of shopping and non-work trips that housewives have to make. Two 2 X 2 contingency tables may be constructed to test this relationship. One pertains to the probability of a respondent stating A being different for men and women, and thus one set of marginal totals is the total number of respondents. This table, of course, corresponds to the first way of calculating percentages mentioned in the preceding paragraph. The other contingency table refers to the number of A responses, by sex, as a fraction of the total number of categories responded to. Here one set of marginals is the total number of categories responded to, and the table as a whole corresponds to the second way of calculating percentages.

Now for the accessibility preference category versus sex of respondent for both contingency tables, df=1. For the first table $\chi^2 = 4.581$, $.02 < p < .05$, while the second is non-significant with $\chi^2 = 2.300$, $.10 < p < .20$. This is not surprising, and indicates that we should be careful as to whether our hypotheses pertain to the fraction of respondents, or the fraction of categories responded to.

The difference in B responses by sex is nonsignificant. For percentage of respondents, the 2 X 2 table has df = 1, $\chi^2 = 1.169$, $.20 < p < .30$.

An overall measure of the relationship between sex of respondent and preferences (all major categories) may be derived by treating the frequencies of Table 7 as a 4 X 2 contingency table, where the marginals are total categories responded to as in Table 5b. As shown at the bottom of Table 7, this relationship is nonsignificant, and the contingency coefficient is quite small.

3.2. Age of Household Head

Age of household head is an important variable, not only in its own right, but as it is associated with the household's stage in the family life cycle, age and number of children, and so on.

We may hypothesize that the percentage of respondents making responses concerning house and lot (category H) is an inverted U-shaped function of age. Households in the middle age range tend to be concerned with the rearing of young children, and are thus oriented towards characteristics of the house and lot, such as house interior space and lot size. 'Older' and (especially) 'younger' households do not have this orientation, but may be said to be more concerned with accessibility and characteristics of the surroundings. For further discussion on family life cycle and residential preferences see, for example, Gans [4].

To test this hypothesis on category H, we may apply to Table 8 the nonparametric Kolmogorov-Smirnov two-sample two-tailed test (Siegel [23, pp. 127-136]), in order to determine

Table 8. Age of Household Head and Residential Preferences

(%/% = % of Respondents/% of Categories Responded to)

Preference Categories	15-24 No.	%/%	25-34 No.	%/%	35-44 No.	%/%	45-54 No.	%/%	55-65 No.	%/%	66 and over No.	%/%	Total No.	%/%
A(Accessibility)	9	60%/27%	42	46%/22%	54	55%/27%	42	51%/28%	22	35%/19%	28	53%/26%	197	49%/25%
H(House and Lot)	10	67%/30%	66	72%/35%	72	73%/35%	53	65%/35%	40	65%/34%	30	57%/28%	271	67%/34%
S(Nat. Env.)	5	33%/15%	36	39%/19%	34	34%/17%	23	28%/15%	20	32%/17%	24	45%/22%	142	35%/18%
N(Non-Nat. Env.)	9	60%/27%	48	52%/24%	43	43%/21%	33	40%/22%	34	55%/29%	27	51%/25%	192	48%/24%
Total No. of Respondents	15	100/	92	100/	99	100/	82	100/	62	100/	53	100/	403	100/
Total No. of Categories Responded to	33	/100	190	/100	203	/100	151	/100	116	/100	109	/100	802	/100

df = 15, χ^2 = 9.373, .80 < p < .90, C = .107

Age of household head unknown for 5 respondents.

whether the respondents stating H were sampled from an age distribution that differs (in any way) from that of the age distribution of those not stating H. For this test, $n_1 = 271$, $n_2 = 132$, the test statistic $D = .1468$, and $.025 < p < .05$, so the hypothesis is accepted.

The same reasoning applied to categories A and M, (Accessibility and Non-Natural Environment) suggests a U-shaped relationship with age of household head. However, inspection of Table 8 does not indicate that much of a relationship is acting, and the Kolmogorov-Smirnov test indicates that the relationships between the age distributions and the A and M responses are both nonsignificant.

In general, Table 8 shows few strong or systematic relationships between preferences and age of household head, as the overall chi-squared test indicates.

3.3. Family Income

Each respondent was asked to state one of the five categories of Table 9 corresponding to the total family income last year, before taxes; 95% responded.

We may hypothesize that the proportion of respondents mentioning H (House and Lot) is negatively related to household income, inasmuch as with increasing income a smaller fraction is generally spent on housing. However, increasing income <u>increases</u> the proportion of respondents mentioning H, contrary to hypothesis, and the Kolmogorov-Smirnov test shows this effect to be statistically significant.

Similarly, we may hypothesize that since a high quality residential natural environment is something of a 'luxury', then

Table 9. Total Family Income (last year) and Residential Preferences

(%/% = % of Respondents/% of Categories Responded to)

Preference Categories	under $5,000 No.	%/%	$5,000-$7,499 No.	%/%	$7,500-$9,999 No.	%/%	$10,000-$14,999 No.	%/%	$15,000 and over No.	%/%	Total No.	%/%
A(Accessibility)	42	59%/30%	50	50%/25%	40	50%/24%	35	50%/23%	28	43%/23%	195	50%/25%
H(House and Lot)	42	59%/30%	56	55%/28%	60	75%/37%	58	83%/39%	45	69%/37%	261	67%/34%
B(Nat. Env.)	25	35%/18%	34	34%/17%	28	35%/17%	30	43%/20%	17	26%/14%	134	35%/17%
N(Non-Nat. Env.)	32	45%/23%	59	58%/30%	36	45%/26%	26	37%/17%	32	49%/26%	185	48%/24%
Total No. of Respondents	71	100/	101	100/	80	100/	70	100/	65	100/	387	100/
Total No. of Categories Responded to	141	/100	199	/100	164	/100	149	/100	122	/100	775	/100

$df = 12, \quad \chi^2 = 13.580, \quad .30 < p < .50, \quad C = .131$

Family income unknown for 21 respondents.

the percentage of respondents mentioning B (the Natural Environment) should increase with family income. No such relationship is found in Table 9, however.

Overall, there are few strong and systematic relations between family income and the preference categories, as indicated by the summary chi-squared statistic.

3.4. Childhood Residence

Sonnenfeld's empirical studies of residential environmental preferences suggest that the physical environment in which a person grew up has a great influence on his residential environmental preferences [24, pp. 71-72]. He emphasizes the necessity of adaptation to an environment different from the type with which the person is familiar by long term contact [24, pp. 74-75].

The questionnaire provides some information with which to assess the role of this variable. Respondents were asked, "During most of the time when you were growing up, did you live in this house, not in this house but in this township, in some other rural area, in some other suburban area, or in another city?" The five possible responses were aggregated to form the three categories of Table 10: rural, suburban, and urban. In making this aggregation, those respondents answering 'other rural', 'other suburban', or 'other city', were assigned the appropriate category of Table 10. In the predominantly rural Brandywine, Neshaminy, and Northern Chester County areas, those answering 'same house' or 'same township were assigned to the 'rural' category. Persons making these two responses in the Main Line area were excluded from categorization. While this assignment, and therefore the design of the childhood residence variable, is undoubtedly crude, it may be useful.

One would hypothesize that since rural, suburban, and urban areas differ most in accessibility and the natural

Table 10. Childhood Residence of Respondent and Residential Preferences

(%/% = % of Respondents/% of Categories Responded to)

Preference Categories	Rural No.	Rural %/%	Suburban No.	Suburban %/%	Urban No.	Urban %/%	Total No.	Total %/%
(Accessibility)	88	47%/23%	36	44%/24%	61	64%/30%	185	51%/25%
(House and lot)	131	70%/35%	56	69%/37%	64	67%/31%	251	69%/34%
(Nat. Environment)	67	36%/18%	23	28%/15%	37	39%/18%	127	35%/17%
(Non-Nat. Env.)	91	48%/24%	37	46%/24%	42	44%/21%	170	47%/23%
Total No. of Respondents	188	100/	81	100/	96	100/	365	100/
Total No. of Categories Responded to	377	/100	152	/100	204	/100	733	/100

$df = 6$, $\chi^2 = 4.630$, $.50 < p < .70$, $C = .079$

Childhood residence of respondent unknown for 43 respondents.

environment, these preference categories would be most affected by the type of childhood residence. Those from urban areas are more likely to mention accessibility considerations than those from rural or suburban areas (64% of respondents compared to 47% and 44%), probably because the former group is more used to high levels of accessibility, and is more sensitive to this attribute when looking for a non-urban residence. This effect is statistically significant, df=2, $\chi^2 = 8.741$, $.01 < p < .02$. However the influence of type of childhood residence on percentage of respondents stating B is nonsignificant, df=2, $\chi^2 = 2.115$, $.30 < p < .50$.

Since no other effects of childhood residence appear in Table 10, it is not surprising that the table as a whole shows nonsignificant relationships.

3.5. Discussion: Interpreting Negative Results

In summary, the preceding analysis has, first, tested certain specific hypotheses between residential preferences and the four socio-economic and demographic variables, and, second, sought to discover whether these variables are 'important' or 'useful' explanations of residential preferences.

In the first case, of the seven hypotheses tested, only two were accepted. There seemed little intuitive explanation of why some hypotheses 'worked' and others didn't; perhaps it was because (see below) very few relationships are apparent among these variables.

In the second case, very rarely were there statistically significant relationships between the preference and the suggested explanatory variables. Out of the 16 possible

relationships between the four explanatory and four preference variables, 3 were statistically significant. For the four whole tables tested, in no case was the relationship between the explanatory variable and the four preference variables statistically significant, nor did it ever even come close.

But statistical significance need not be the only criterion. Inspection of Tables 7 to 10 suggests that there appear to be very few strong, systematic, simple relationships. Thus, the percentage of respondents stating a given preference category does not vary very much across a row, and the variation that does occur may have little obvious interpretation.

A possible reason for these negative results is that the relationships are undoubtedly more complex than those assumed by this single-explanatory-variable-at-a-time mode of analysis. Perhaps all of the explanatory variables used here (and others) should be used at once in a multiple regression or analysis of variance approach.

Furthermore, it is likely that there is considerable 'noise' or imperfect association in the relationship between the theoretical preference concepts of interest and the actual preference data used. This imperfect association may be so high as to make our variables very poor indicators of residential preferences, by 'submerging' any relationships that may be acting.

Now the problems of possibly complex relationships and imprecise measures of the theoretical variables of interest are all common troubles in exploratory empirical research. These problems make the interpretation of negative results

a difficult and perhaps bootless task, and are one reason why negative results are rarely published. The problems make it difficult to state that negative results prove the lack of a relationship.

The justification here, though, for the presentation and discussion of the negative results is the fact that these preference data have yielded positive and substantively interesting results when analyzed differently, as was seen in Section 2 above and will be shown in Section 5, below. The presence of some positive results using the same preference variables and relationships as simple as in this section may indicate that the data have some 'qualit and the relationships that are acting have some simplicity.

4. RESPONDENTS' RESIDENTIAL CHOICE

In order to compare respondents' residential preferences with their actual choice of a place to live, data were collected on each respondent's residence, using variables paralleling the preference variables. To do this, some of the preference categories were narrowed, and there was a necessity to obtain difficult numbers, such as quantitative measures of the quality of the natural environment. This section discusses the residential choice variables used. Each residential choice variable is denoted by a symbol of the form q_i, indicating the quantity or level of residential attribute i. The variables, their means and standard deviations are summarized in Table 11.[8]

Accessibility to work, q_w. The general preference category of accessibility, A, was narrowed to just accessibility to

---

[8]To obtain as great a diversity in residential environments as possible, the 44 respondents in the Towns area were included in the analysis of this and the following sections, bringing the number of respondents to 457.

TABLE 11

Means and Standard Deviations for Residential Choice Variables (see text)

Symbol	Name	Mean	Standard Deviation	No. of Observations
Accessibility				
q_w	$-\log_{10}$ (commuting time)	-1.286	0.366	337
Lot Size				
q_ℓ	$\log_{10}$(lot size)	0.241	0.878	416
Natural Environment				
$q_{b,1}$	stream, 1/2 mi.	2.494	1.995	92
$q_{b,2}$	stream, 1 mi.	9.708	4.930	92
$q_{b,3}$	pond, 1/2 mi.	0.236	0.454	92
$q_{b,4}$	pond, 1 mi.	1.030	1.149	92
$q_{b,5}$	local relief, 1/2 mi.	184.522	84.910	92
$q_{b,6}$	local relief, 1 mi.	374.663	82.787	92
$q_{b,7}$	forest, 1/2 mi.	20.946	16.396	92
$q_{b,8}$	forest, 1 mi.	21.002	13.722	92
$q_{b,9}$	beauty rating	9.370	2.484	92
Non-Natural Environment				
$q_{c,1}$	-(no. of structures), 1/2 mi.	-139.033	236.404	92
$q_{c,2}$	-(no. of structures), 1 mi.	-430.380	589.170	92
$q_{c,3}$	-traffic, 1/2 mi.	-4,648.	6,420.	92
$q_{c,4}$	- traffic, 1 mi.	-15,825.	16,980.	92

work, so that the choice variable q_w was defined in terms of the length of the journey to work, in minutes, for the household head. This information was obtained for 337 (or 74%) of the 457 respondents. For the 120 others, either the household head did not work, worked at home or at no particular location, or the respondent did not know, or gave no answer. The actual variable used was $-\log_{10}$ (commuting time), where commuting time was measured in minutes. The logarithm of commuting time was taken because it was felt that equal ratios rather than equal differences of commuting times were perceived as equal; that is, a 60 minute worktrip stood in the same relationship to a 30 minute worktrip, as a 30 minute worktrip did to a 15 minute one. Taking the negative of the logarithm produced a measure of accessibility (rather than inaccessibility), comparable to the preference for accessibility.

Lot size, q_ℓ. The major preference category of characteristics of the house and lot, H, was narrowed down to just lot size, which was compared to the respondent's actual lot size. Of the 457 respondents, 416 (or 91%) stated their lot size. The remaining individuals either lived in apartments, did not know their lot size, or did not supply this information. The actual variable used was $\log_{10}$ (lot size) where lot size was measured in acres; the logarithmic measure was used for the same reasons as given above for accessibility.

The following 13 variables pertain to the natural and nonnatural environment. While the data above were taken from the questionnaires, the data below were obtained from maps, aerial photographs, and field surveys. Collecting the

data for the environment of each of the 457 respondents would have been prohibitively expensive, so respondents were aggregated spatially. Environmental data were collected for each of the 92 segments, the clusters used in the sampling design, containing an average of 5.0 respondents. The segments are spatially compact and generally rectangular. Outside the towns area they are bordered by roads and range from about a quarter to one square mile in area. After the data were collected for each segment, those data were assigned to each of the respondents living in that segment, effectively 'enlarging' the sample from 92 observations to 457. However, this 'enlargement' was done for statistical convenience and the number of independently sampled observations remains 92. The observations on the 92 original segments were used for the computations of Tables 11 and 12, the latter table being a correlation matrix for the 13 environmental choice variables.

Another problem in the design of the environmental choice variables was the definition of the area over which they were to be measured. This question is a research topic in itself, requiring a quantitative knowledge of persons' perception of the natural environment. For simplicity, areal units were taken that were independent of the exact size and shape of the segment, on the assumption that the area relevant to perception was relatively constant across segments. Furthermore, in order to experiment, two differently sized areal units were used for all variables except $q_{b,9}$. The segment's population center of gravity was located approximately, using USGS topographic maps, 1:24,000 scale. One areal unit was a 1/2 mile radius circle, the other unit was a circle of a mile radius; the center of both circles was the segment

population center of gravity. Thus (referring to Table 11) $q_{b,1}$ is the stream variable for the 1/2 mile circle, and $q_{b,2}$ the stream variable for the 1 mile circle.[8a]

The first nine variables below, indicated by the subscript 'b', are intended as measures of the quality of the respondents' natural environment. This is admittedly a difficult quantity to measure, so the variables represent five different characteristics of the natural environment. The analysis below either treats each of the 9 variables individually, or averages statistics derived from them.

Stream, $q_{b,1}$, $q_{b,2}$. This variable is defined as the total length of streams in each area (1/2 or 1 mile circle), since streams presumably contribute to an area's natural beauty and one measure of their 'presence' in an area is total stream length. Stream length was measured in arbitrary units from the USGS 1:24,000 topographic series. Rivers were differentiated from streams by defining the former to be wide enough so that the blue line on a map had a 'filled in' appearance. Data were collected on river length also, but it was found that in the study area rivers

---

[8a]An alternative approach to areal unit definition would be to use the stream sub-basin for each respondent's residence. This is practical, and seems a 'natural' areal unit both for physical environmental characteristics, and in terms of persons' environmental perception, the latter because the ridge lines defining the sub-basin are associated with one's line of sight. Furthermore, the use of sub-basins, rather than the areal units adopted here, should materially increase the correlations among the physical environmental variables $q_{b,1}$ to $q_{b,8}$; see below and Tables 11 and 12. I am indebted to James C. Knox for this suggestion.

were the site of concentrations of people, manufacturing, and sometimes water pollution, so that river length was not a good measure of natural environmental beauty.

Pond, $q_{b,3}$, $q_{b,4}$. The total perimeter of ponds and lakes in each circle was measured from the topographic maps, using arbitrary length units. Again, ponds were thought to be one element in the area's beauty, and, since one thinks of a pleasant location as being along a lake's periphery, the total pond perimeter (not area) was thought to be a good measure of the ponds' visual impact on the area.

Local relief, $q_{b,5}$, $q_{b,6}$. Local relief is the maximum difference in elevation in each area that is, the difference between the highest and lowest points in each circle. It is measured in feet, from the USGS topographic maps. Local relief is a measure of hilliness, or the roughness of the area's topography, which presumably is one element in an area's natural beauty.

Forest, $q_{b,7}$, $q_{b,8}$. The percentage of each area that is forested was measured from a 1965 aerial photographic survey, using prints at a 1" = 800' scale.

Beauty Rating, $q_{b,9}$. A scale was developed as a quantitative measure of people's evaluation of natural environmental beauty (Coughlin and Goldstein [2]). Two judges were taken to each site and asked to rate the quality of the natural environment on a one to seven scale, with seven indicating the highest level. The judges were instructed to make the difference between ratings two and three indicative of approximately the same difference in environmental quality as that between ratings five and six. Thus, the scales could

be treated as interval variables, rather than just rankings. The sum of the judges' ratings for each segment constituted $q_{b,9}$. Study of these scales indicates a high level of inter-judge agreement; for more discussion see Coughlin and Goldstein [2] and Rabinowitz and Coughlin [21].

The remaining four variables, designated by the subscript 'c', measure the density or congestion of the nonnatural environment. The major preference category M, characteristics of the nonnatural or man-made environment, was narrowed to just a preference for the avoidance of many people in the vicinity, response type 5 in Table 2. The negative of the quantities below was taken, since the preference variable pertains to avoidance of congestion.

Number of Structures, $q_{c,1}$, $q_{c,2}$. A feasible measure of population density was the total number of structures in each circle. Since all areal units were one of two sizes, the number of structures was used, rather than structures per unit area. This variable was measured from the aerial photographs used for $q_{b,7}$ and $q_{b,8}$.

Traffic, $q_{c,3}$, $q_{c,4}$. Another indicator of the presence of human activity was taken to be the traffic flow through each area. Using county planning commission data, the traffic flow (e.g., average number of vehicles per day), for each road in each area was identified. This flow was multiplied by the length of the portion of each road in a given circle, (in arbitrary units) to obtain that road's flow mileage in the circle. Flow mileage was summed across all the roads in that circle to obtain the circle's total traffic measure. Flow mileage was thought to be an indication of the area's

perceived traffic since it encompasses both traffic flow and the length of roads in the area, the 'presence' of that road in the area.[9]

Tables 11 and 12 allow a brief investigation of the statistical properties of environmental variables. Standard deviations are generally large relative to the mean level, indicating that the areas do in fact differ in these indicators of residential environments.

The diversity of physical environments, and the fact that the various $q_{b,i}$ variables measure quite different environmental characteristics, are suggested explanations for the low intercorrelations among the $q_{b,i}$ variables in Table 12a. The 36 correlation coefficients have an average of just .206. Even the correlations between the same variables defined on half and whole mile circles is relatively low, in spite of the fact that the former circle of course constitutes a quarter of the latter's area. These four 'part-whole' correlations have an average r of .559. The remaining 32 coefficients among different variables average .162. While these correlations are small, they are predominantly positive, suggesting that the $q_{b,i}$ variables are measuring similar or related environmental characteristics. Only two correlation coefficients are negative, and these are small in absolute value. While only 13 of the 36 coefficients are statistically significantly positive taken individually,

---

[9]The Pennsylvania Turnpike ran through several of the circles, but none of the turnpike's interchanges fell inside the segments. Because this is a limited access highway, without bordering residences, the road's large flow was excluded from the flow mileage calculations.

Table 12. Product Moment Correlations Among Residential Environmental Variables (see text)

a. Correlations Among Individual Variables

		Natural Environment									Non-Natural Environment			
		$q_{b,1}$	$q_{b,2}$	$q_{b,3}$	$q_{b,4}$	$q_{b,5}$	$q_{b,6}$	$q_{b,7}$	$q_{b,8}$	$q_{b,9}$	$q_{c,1}$	$q_{c,2}$	$q_{c,3}$	$q_{c,4}$
Natural Environment	$q_{b,1}$	1.000	.732	.154	.142	.144	.077	.270	.173	.180	.275	.184	.280	.064
	$q_{b,2}$		1.000	.144	.165	.217	.039	.244	.311	.117	.265	.253	.215	.065
	$q_{b,3}$			1.000	.499	-.016	-.022	.185	.041	.092	.002	-.046	-.176	-.228
	$q_{b,4}$				1.000	.009	.026	.167	.036	.137	-.097	-.116	-.260	-.326
	$q_{b,5}$					1.000	.430	.601	.351	.148	.192	.218	.293	.172
	$q_{b,6}$						1.000	.274	.259	.101	.229	.194	.307	.199
	$q_{b,7}$							1.000	.575	.250	.281	.229	.331	.176
	$q_{b,8}$								1.000	.181	.116	.193	.103	.088
	$q_{b,9}$									1.000	.342	.289	.380	.328
Non-Natural Environment	$q_{c,1}$										1.000	.822	.751	.678
	$q_{c,2}$											1.000	.662	.644
	$q_{c,3}$												1.000	.855
	$q_{c,4}$													1.000

Two-tailed significance levels N = 92, p = .05, |r| = .205
p = .01, |r| = .267

b. Average Correlation among Sets of Variables

	$\bar{q}_b$	$\bar{q}_c$
$\bar{q}_b$	.206	.139
$\bar{q}_c$		.735

($p = .05$, two-tailed test) the overwhelmingly positive overall pattern strongly suggests that the $q_{b,i}$ variables <u>taken as a whole</u> are significantly positively correlated.

The four $q_{c,i}$ variables, on the other hand, are more strongly positively intercorrelated. This is not surprising since for most of the sample areas the traffic is largely non-commercial and thus would be strongly associated with population density, and the roads serve the homes, so their mileage would also be associated with population density. The overall level of intercorrelation is .735; the two 'part-whole' r's average .839, while the four others average .684.

The off-diagonal block in Table 12a indicates that the $q_{b,i}$ variables are, on the whole, positively associated with the $q_{c,i}$ variables. Thus, of the 36 correlation coefficients, 7 are negative, and 3 of these obtain two-tailed statistical significance for $p = .05$. All of the negative correlations appear for the two pond variables, $q_{b,3}$ and $q_{b,4}$. Of the 29 remaining coefficients, 16 are statistically significantly positive.

The positive correlations in this off-diagonal block indicate simply that by our measures, areas that have high quality natural environments are likely to have low levels of congestion, not a surprising result. Also of interest is the <u>level</u> of correlation. The average of the $q_{b,i}$, $q_{c,j}$ correlation coefficients is .139 (including the negative coefficients) or .216, excluding the negative coefficients. These levels are quite similar to the average intercorrelation, .162, among the $q_{b,i}$ variables, when

the 'part-whole' correlations are excluded. Thus the natural environment variables are just about as strongly associated with lack of congestion, as they are with each other.

5. RELATIONSHIPS AMONG RESIDENTIAL PREFERENCES AND CHOICE

In this section the residential choice variables of the previous section are related to measures of residential preferences in order to see whether our preference concepts, categories, and variables are usefully linked with actual residential choice behavior. The symbol p_i indicates the level of preference for residential attribute i. First, the residential preference variables of Sections 2 and 3 are modified somewhat.

5.1. Definitions of Relationships Among the Preference Variables Used

The respondent's preference for the natural environment, p_b, is defined as the number of times response types included in major preference category B (see Table 2) were mentioned, divided by the total number of responses made by that respondent. For the three other preference variables, response frequencies were also divided by the total number of responses and the definition of the preference variable was narrower than for the major preference categories. Preference for accessibility to work, p_w, a part of major category A, is calculated from the frequency of response type 14 in Table 2. Preference for a large lot, p_l, part of major category H, is calculated from the frequency of response type 23. Preference for the avoidance of densely populated or congested surroundings, p_c, is calculated from the frequency of response type 5. (See discussion in

Section 2 of the use of the response types to measure density preferences.) Distributions of response frequencies for the last two variables (not divided by the total number of responses) appear in Table 3.

Table 13 presents some of the statistical properties of the four preference variables. [There were no preference data for 8 respondents.] Observe that several of the response types were mentioned quite seldom, so that the associated preference variable has very few positive values. Work accessibility and the avoidance of congestion were mentioned at all by only 14% and 10% of the respondents. Very intuitively, our sample is much 'smaller' and has much less 'variation' than may be apparent, because 'positive' data on these variables have been obtained from few persons. This problem is one possible explanation of the extremely low correlation coefficients that bedevil this section's analysis. However, the lower response frequencies are in part due to the decision to have preference variables that are directly comparable to the choice variables, as well as to the unstructured nature of the residential preference question.

Let us first consider relationships among the preference variables by computing product-moment correlation coefficients of the form $r(p_i, p_j)$ where i and j run over all the preference categories. These coefficients, computed from a sample of 449 with no missing observations, appear as the upper left block in Table 16. This submatrix indicates preference relationships similar to Table 6b, in spite of the narrowed preference categories, continuous as opposed to binary variables, and

Table 13. Residential Preference Variables: Percentage of Nonzero Responses, Means and Standard Deviations, N = 449

Symbol	Name	% of Positive Responses	Mean	Standard Deviation
P_w	accessibility to work	14.0%	0.046	0.139
P_l	lot size	20.9	0.058	0.132
P_b	natural environment	39.0	0.120	0.208
P_c	avoidance of congestion, density	9.8	0.030	0.097

product-moment versus contingency correlation coefficients. All the preference variables are either uncorrelated or negatively correlated, except preferences for the natural environment and the avoidance of congestion which are positively associated. However the coefficients are quite small in magnitude in Table 16 and barely attain statistical significance, in spite of the large sample size.

5.2. Relationships Among the Choice Variables

Similarly consider relationships among the residential choice variables, the lower right submatrix in Table 16. Some of the individuals used in computing the whole correlation matrix had data missing for q_w, q_ℓ, or both (see bottom of Table 16); in computing the correlation matrix these individuals were assigned the mean value of q_w or q_ℓ. This way of treating missing observations does not tend to produce a relationship when none exists, but has the effect of 'diluting' any existing relationship, reducing the magnitude of the correlation coefficients. As in Table 12b, the symbols $\bar{q}_b$ and $\bar{q}_c$ indicate that those entries are averages of the correlation coefficients for the individual $q_{b,i}$ and $q_{c,i}$ variables. These averages are different from those in Table 12b, for those correlations were computed over the 92 segments; the ones in Table 16 are computed over the 449 respondents. Also observe that it is hard to say what the relevant sample size is for testing the statistical significance of a correlation where at least one of the variables is a $q_{b,i}$ or $q_{c,i}$, since these data were

obtained from but 92 independently sampled observations. Thus, the bottom of Table 16 lists significance levels of r for both N = 449 and 92. Below, whenever a correlation containing at least one $q_{b,i}$ or $q_{c,i}$ is said to be statistically significant, that statement is conservative, for N = 92 is used in the computations.

Although the magnitudes of the q_i q_j correlation coefficients are small, their signs are indicative of what might be expected for the metropolitan area as a whole. Thus, the negative $r(q_w, q_\ell)$ indicates the negative association, across various residences, of accessibility to work and lot size; roughly, the more accessible residences have smaller lots. Similarly, the remaining correlations indicate that more accessible residences have lower quality natural environments; that more accessible residences are in more densely populated areas; that residences with large lot sizes are located in areas with high quality natural environments (though the magnitude of this r is vanishingly small); that large lots are, not surprisingly, located in sparsely populated areas; and that natural environmental quality is associated with sparse population. That these relationships hold for the sample areas is at least a bit surprising: the sample completely excludes central city and inlying suburban locations; commuting patterns of outlying satellite cities (the Towns area) would seem to be representative of neither city nor suburb; and in general the sample is hardly representative of the combination of these variables for the metropolitan area as a whole. However these relationships are

gratifying, for they do help justify any inferences made from this sample to the metropolitan area as a whole.

5.3. Some Theory About First- and Second-Order Preference-Choice Relationships

Having discussed our preference and choice variables, we may proceed to the meat of this section: relating residential choice to residential preferences. The basic hypothesis here is that, all other things being equal, preferences express themselves through choice. Thus, we would expect that all else constant a person with a high preference for a large lot or a beautiful natural environment would be more likely to have a residence with a large lot or a beautiful natural environment than someone without these preferences. To work with these concepts we must discuss the 'all things being equal' caveat above, as well as the meaning of 'preferences manifesting themselves through choice'.

Clearly, if the residential preference concept is useful there must be some relationship between preferences and choice. Furtner, if our operational measures of preference and choice are also of some use, there will be some relationship between the two sets of variables. In both cases we speak of some relationship or talk of holding other things equal, because influences other than preferences are surely acting on residential choice. The other influences include income, information as to housing opportunities, changes in preferences and the characteristics of the residence since the time of the move, and uncertainty in general. High income may make it possible to

afford high levels of several desirable residential attributes, rather than just the most preferred. Given two persons with the same preferences, it is reasonable to expect the richer to have, say, a larger lot than the poorer. One would expect the level of information about housing opportunities to act similarly to income: well-informed persons would tend to have more desirable choices, for all residential attributes, than poorer informed persons with the same preferences and income.

One obvious and important reason for a 'mismatch' between preferences and choice is changes since the move that are insufficiently important to cause the individual to seek a new location. Assume that at the time of the move the residence matches one's preferences perfectly. In time, residential preferences may change, as children grow older and the need (and therefore preference) for a large lot diminishes. Alternately, characteristics of the residence may themselves change, as if the surrounding area becomes subdivided. Finally there may be uncertainty about attributes of the residence, or about one's own preferences: "We didn't know we would miss the convenience of the city so much."

Let us now consider two different ways of relating preferences and choice. One way might be termed the <u>first-order</u> relationship between <u>absolute</u> levels of the preference and choice variables. Thus, assuming a linear relationship, one would calculate product-moment correlation coefficients of the form $r(p_i, q_i)$ where i runs over w, ℓ, b, and c. If preferences do manifest themselves through choice, and the other influences

on choice have no systematic effect (that is have a zero expectation so they 'cancel out' on the average) then this correlation coefficient should be positive.

An intuitive interpretation of this preference-choice relationship may be made from the microeconomic theory of consumer behavior (see, e.g., Henderson and Quandt [7, Ch. 2]). Consider the four residential attributes as 'commodities', q_i as the quantity of each residential commodity that is being consumed, and the utility function as defined on the four q_i's and q_z, the quantity of all other goods consumed. When utility is maximized the consumption of each commodity is increased until the marginal utility of it is equal to its price times a quantity often interpreted as the marginal utility of money.[10] Now in the present analysis p_i may be interpreted as the marginal utility of i, given some bundle of all the goods consumed. Very intuitively, the larger this is, the higher q_i can be before its marginal utility is equated to its price times the marginal utility of money, so, all things being equal, there should be some monotonic relationship between p_i and q_i.[11]

---

[10]The latter quantity is the partial derivative of the Lagrangean expression with respect to income.

[11]Making highly intuitive interpretations of precise formulations like the Theory of Consumer Behavior is a dangerous enterprise, because the assumptions implicit in the sentence above are not at all obvious. I am currently working on deducing a well-defined relationship between p_i and q_i from consumption theory, as well as deducing the form of the second type of preference-choice relationship, discussed below. Several encouraging results have already been obtained.

The other way to relate preferences and choice considers the second-order relationships between relative levels of preferences and choice. Thus, given accessibility to work (w) and quality of the natural environment (b) as both desirable, a person with a high p_w relative to his p_b should have a residence with a high q_w relative to its q_b. Again under the assumption of a linear relationship, we would expect the product-moment correlation of $r((p_w-p_b), (q_w-q_b))$ or, more generally, of the form

$$r((p_i-p_j), (q_i-q_j)) \qquad i \neq j;\ i, j = w, \ell, b, c.$$

to be positive.

The second-order preference-choice relation can also be interpreted in the light of consumption theory. As is well known, at individual equilibrium the ratio of the marginal utilities of any two commodities, or the rate of substitution among them, is proportional to the inverse of the price ratio of the two commodities. This price is assumed the same for all individuals. Additionally assume, as above, that the larger p_i or p_j is, the greater q_i or q_j will be at the point where the rate of commodity substitution equals the price ratio. The conversion of the preference-choice relationship from one defined on ratios to one defined on differences is accomplished by taking logarithms. See the definitions of q_w and q_ℓ, Section 4.

Now, there are two reasons why (at least to the author) the second-order preference-choice relationship is intuitively

more appealing than the first-order one. For one thing, the very notion of preferences, as used here, involves the concept of trading off among several residential attributes, each of which is itself desirable. Preference, then, involves relative desirability, which may plausibly be expressed by relative levels of the attributes at the individual's residence.

Second, using relative quantities would appear to cancel out, or at least reduce the importance of, other influences on choice such as those stated above. Thus, while high levels of income or information would tend to raise the levels of both q_i and q_j, whatever p_i or p_j, they would seem to have much less of an effect on the relative levels of q_i and q_j. This point can be made explicit in the case of income, with the example from consumption theory. In the first-order preference-choice relationship income influences q_i because it influences the marginal utility of money. Under the usual utility function assumptions, when an individual's income is increased, the marginal utility of money diminishes and, ceteris paribus, the marginal utility of q_i attains equality with the product of the price of i and the marginal utility of money, at a higher q_i. Otherwise expressed, for two individuals with identical preferences (identical utility functions), facing the same prices, the individual with the higher income will have a higher level of q_i.[12]

---

[12] We are of course assuming that none of the residential attributes is an inferior good, that is, a commodity whose consumption is lowered by increased income.

Whatever the appeal of a second-order preference-choice relationship, its use here suffers from the major problem that the correlation coefficient computed is dependent on the inherently arbitrary scale units used for q_i and q_j. This is because the quantity $(q_i - q_j)$ is influenced by the scales used to measure q_i and q_j. For example, the quantity (temperature-pressure) is influenced by whether temperature is measured in Centigrade or Farenheit. In consumption theory, q_i may be measured by the units of that commodity, or dollars spent on the commodity. Both scales are 'natural' and 'comparable' across different commodities. In the case of the former scale, measuring q_i in units or dozens of units consumed of course changes q_i, but this change also affects the price proportionately, so results are invariant with regard to units.

In the present case, it is unlikely that we have 'natural' units for q_b, much less ones that are comparable to the other q_i's. Indeed, the units of the component variables that make up q_b are themselves neither natural nor comparable, and it is even hard to name them. Until it can be shown that the result of any analysis is scale invariant, or that the scales used are somehow 'natural', it would seem wise to avoid the use of quantities like $(q_i - q_j)$.[13] Thus, let us initially concentrate attention on first order preferences.

No such scaling problems are associated with correlation coefficients of the form $r(p_i, q_i)$ since, as is well known,

---

[13] It is shown below that under certain assumptions about first order preferences, the sign of $r((p_i-p_j), (q_i-q_j))$ is scale invariant.

product-moment correlations are invariant with positive linear transformations of the data.[14] But we could also measure association between p_i and q_i with a rank order correlation coefficient, therefore avoiding any assumption of a linear relationship. The value of the rank order correlation coefficient is of course invariant with any monotonic (order-preserving) transformation of the data.[15]

There is another reason for the initial use of first- rather than second-order preference-choice measures. This is that the predicted sign of $r((p_i-p_j), (q_i-q_j))$ may not be clear for all pairs i, j. Given the relation between preferences for a beautiful natural environment and the avoidance of high density or congestion, what would the sign of $r((p_b-p_c), (q_b-q_c))$ be?[16]

5.4. First-Order Preference-Choice Relations

Now, in the first order choice-preference relationship, we may also be interested in correlations of the form $r(p_i, q_j)$, $i \neq j$. For example, $r(p_b, q_\ell)$ indicates the relationship between natural environmental preferences and lot size

---

[14] That is, defining the new variable $q'_i = \alpha q_i + \beta$, then for all β and all α positive, $r(p_i, q_i) = r(p_i, q'_i)$.

[15] That product-moment rather than rank order correlation coefficients are used below is due partly to computational convenience and partly to the fact that when the assumptions of the former are justified, they are a more efficient measure of correlation than the latter (Siegel [23, p. 223]). With the low correlation levels to be shown, high efficiency for correlation measures would seem to be quite important.

[16] As is indicated below, Table 17a shows this fear to be justified.

chosen. It is fortunate that four simple, intuitively appealing axioms or assumptions can be used to generate the signs of all sixteen correlation coefficients of the form $r(p_i, q_j)$, for all combinations of i and j. In other words, these axioms generate the qualitative form (the signs of the elements) of the entire off-diagonal block of the correlation matrix of Table 16.

The axioms are listed in Table 14 and Table 15 shows the predicted qualitative correlation matrix deduced from them. Axiom I is obvious: preferences manifest themselves in choice, as has been stated. Axiom II indicates that since preferences for the natural environment and the avoidance of congestion are 'similar', $r(p_b, q_c)$ and $r(p_c, q_b)$ should have the same positive sign as $r(p_b, q_b)$ and $r(p_c, q_c)$.

The third axiom was suggested by interviews with residents of rural areas. Persons preferring a pretty natural environment or low density, and who are living in an area that presently has these attributes, may nevertheless fear changes in their surroundings with time. The forest across the road may become a subdivision or a junkyard. The municipality may have no land use controls strong enough to prevent this.[17] This fear may cause the homeowner to buy a larger quantity of land than might otherwise be the case. He does this to "protect his view" or "protect his frontage". Thus, a preference for b or c would tend to manifest itself in a high q_ℓ, and result in the correlation coefficients $r(p_b, q_\ell)$ and $r(p_c, q_\ell)$ being positive.

---

[17] Evelyn Waugh has written a short story, set in rural England, in which this fear is exploited by a clever outsider; see [25].

Table 14. Axioms Generating Qualitative Correlation Matrix Between Preference and Choice Variables, Table 15 and Off-Diagonal Block, Table 16

I (Expression of Preferences in Choice)

$$r(p_i, q_i) > 0, \qquad i = w, \ell, b, c.$$

II (Similarity of Preferences for Natural Environment, Low Density)

$$r(p_b, q_c) > 0, \qquad r(p_c, q_b) > 0.$$

III. ("Protecting my environment;" High p_b, p_c manifested by large lot)

$$r(p_b, q_\ell) > 0, \qquad r(p_c, q_\ell) > 0.$$

IV. (Substitutability of residential attributes)

<u>except for II and III</u>, $r(p_i, q_j) < 0, \qquad i \neq j,$

$$i, j = w, \ell, b, c.$$

<u>e.g.</u>, $r(p_w, q_\ell) < 0.$

Table 15. Predicted Qualitative Preference-Choice Correlation Matrix (see text)

Preference Variables	Choice Variables q_w	q_ℓ	q_b	q_c
P_w	$+^{**}$	$-^{**}$	$-^{*}$	$-$
P_ℓ	$-^{*}$	$+^{**}$	$-^{*}$	$-$
P_b	$-^{*}$	$+^{**}$	$+^{*}$	$+^{**}$
P_c	$-^{*}$	$+^{*}$	$+^{*}$	$+^{*}$

*correct sign predicted, see Table 16.

**correct sign predicted and correlation coefficient statistically significant, one-tailed $p \leq .05$, see text for relevant sample size, and Table 16.

When many people manifest high levels of p_b or p_c by large lots, this tends to bring about a relatively low density urban sprawl development pattern. Following this reasoning, given effective land use controls for environmental preservation, persons with the same residential preferences would choose smaller lots, and this cause of urban sprawl would be reduced.

Notice that our reasoning does <u>not</u> support what may be called the 'converse relationship': that $r(p_\ell, q_b)$ and $r(p_\ell, q_c)$ be positive. A desire for a large lot need not manifest itself by a pretty natural environment. Indeed Table 16 shows that at least the first of these correlations is not positive but negative.

The fourth axiom indicates that (with exceptions as stated in axioms II and III) the four residential attributes serve as 'substitute commodities' for each other. That is, if i and j are substitutes, a high level of p_i is associated with a low level of q_j, since this means that the respondent is interested in i rather than j. One example is that $r(p_w, q_\ell)$ is predicted to be negative: an interest in high accessibility to work would lead, other things being equal, to a smaller lot size.

Table 15 compares the qualitative predictions with the actual correlation coefficients, from Table 16. Out of the 16 correlation coefficients, the signs of 14 are predicted correctly. The two coefficients whose signs were predicted incorrectly, $r(p_w, q_c)$ and $r(p_\ell, q_c)$ are quite small. At .019 and .010 these coefficients are the third and first smallest

Table 16. Product Moment Correlations of Residential Preference and Choice Variables, N = 449, Mean Value of that Variable Inserted for Missing Observations (see text)

		Preference Variables				Choice Variables			
		P_w	P_ℓ	P_b	P_c	q_w	q_ℓ	$\bar{q}_b$	$\bar{q}_c$
Preference Variables	P_w	1.000	-.076	-.098	-.048	.133	-.121	-.057	.019
	P_ℓ		1.000	-.146	-.029	-.058	.078	-.015	.010
	P_b			1.000	.123	-.043	.175	.057	.226
	P_c				1.000	-.026	.069	.031	.137
Choice Variables	q_w					1.000	-.121	-.054	-.140
	q_ℓ						1.000	.015	.270
	$\bar{q}_b$							.195	.049
	$\bar{q}_c$								.693

One-Tailed Significance Levels
(double probabilities for two-tailed levels)

N = 449, p = .05, |r| = .078,
p = .025, |r| = .092,
p = .01, |r| = .109,

N = 92, p = .05, |r| = .173,
p = .025, |r| = .205,
p = .01, |r| = .242.

Numbers of Missing Observations for Correlation Matrix Above

q_w	120
q_ℓ	41
q_w and q_ℓ	14

(in absolute value) in the entire $p_i q_j$ matrix, so it is not surprising that these predictions should be in error in a qualitative sense.

Even though the vast majority of correlation coefficients had their signs predicted correctly, most coefficients are quite small in magnitude. The average absolute value of the coefficients is only 0.078. The small size of the coefficients causes statistical significance to be obtained in but 5 of the 14 coefficients whose signs were predicted correctly.[18]

The small size of the correlation coefficients is not surprising, considering all the practical and theoretical problems involved in the use of these data. As has been mentioned, our measures of preference and choice are undoubtedly imperfect and thus 'noisily' and imperfectly associated with the true theoretical preference and choice variables. Furthermore, statements concerning first-order preference-choice relationships require strong ceteris paribus assumptions. As is well known, when these

---

[18] $p \leq .05$, one-tailed, since the sign of the correlation coefficient was specified by the alternative hypothesis. See Section 5.2 concerning the sample size used to calculate the significance levels. As before, $\bar{q}_b$ and $\bar{q}_c$ indicate that the correlation coefficients in that row or column are the averages of the 9 or 4 correlations computed using the individual $q_{b,i}$ and $q_{c,i}$ variables; see Section 4.

Using the 'raw' correlations for the $q_{b,i}$ and $q_{c,j}$ variables, one may construct a 4 X 15 correlation matrix replacing the 4 X 4 matrix of Table 15. For the 60 correlation coefficients in the former matrix, 48 have signs correctly predicted by the four axioms, where the reasoning applied to correlations including q_b applies to each of the $q_{b,i}$ variables. Of the 48 coefficients with correctly predicted signs, 9 of them are statistically significant, according to the criterion above.

'other things' are not equal, but do vary in the sample, then the magnitude of the correlation is likely to be reduced.

What is surprising is the close correspondence between the complex, non-intuitive pattern of the predicted qualitative correlation matrix, and the signs of the observed correlation matrix. This situation is reminiscent of Dr. Samuel Johnson's celebrated Talking Dog. The thing about the Talking Dog was not how well it talked, but that it talked at all. Furthermore, a series of significance tests on individual correlation coefficients is not the appropriate test of our hypothesis. Our alternative hypothesis specifies a particular qualitative form of the whole off-diagonal block of the correlation matrix, and the appropriate test should be one on this whole matrix. Significance levels of individual correlation coefficients bear no simple relationship to the outcome of a test on the whole matrix because, among other reasons, the intercorrelations among the variables make the sample correlation coefficients themselves distributed dependently.

Among standard procedures, the one coming closest to testing our hypothesis is the 'Test for Complete Independence', which is related to the canonical correlation procedure (see Morrison [16, pp. 207-213]). This test has as its null hypothesis that the population value of the off-diagonal block of a partitioned correlation matrix (like Table 16) is a null matrix, in other words, $r(p_i, q_j) = 0$ for all i, j. The alternative hypothesis is that this matrix is non-null, that is, that there is at least one i, j pair so that $r(p_i, q_j) \neq 0$. This alternative hypothesis is far more general than is required in the present

case, since Table 15 specifies the sign of all correlation coefficients, and all coefficients are nonzero. Work is now underway on developing a Generalized Likelihood Ratio Test for the specific alternative hypothesis represented by Table 15.

It is possible, though, to develop a simple test for the statistical significance of the whole $r(p_i, q_j)$ correlation matrix. One possible null hypothesis is that the population correlation coefficient, $\rho(p_i, q_j) = 0$, for all i, j, so that the population correlation matrix is null. This hypothesis simply posits that there is no relationship between the set of p variables and the set of q variables. But our alternative hypothesis (Table 15) states not only that there is some relationship between the two sets, but also specifies the sign of all sixteen correlation coefficients.

A possibly more appropriate null hypothesis is that $\rho(p_i, q_j) \leq 0$, for those i, j pairs predicted as positive by Table 15, and $\rho(p_i, q_j) \geq 0$ for those i, j pairs predicted as negative by Table 15. Under the more general null hypothesis, the parameter subspace (space of all $\rho(p_i, q_j)$ values consistent with the hypothesis) is the complement of the subspace under the alternative hypothesis. However this null hypothesis would seem quite difficult to test by the usual means. It specifies a whole region (actually a half-space) in the parameter space, rather than a point, as is usual. Thus the joint sampling distribution of the $r(p_i, q_j)$'s under the null hypothesis is not one multivariate distribution but a whole family of distributions, corresponding to different 'versions' of the null

hypothesis. For such a null hypothesis we cannot calculate a unique number for the statistical significance of the sample, that is, the probability that the observed data were sampled under the null hypothesis.

We can, however, make a conservative statement as to the statistical significance of the data. As an example take an individual correlation coefficient where the initial null hypothesis is that $\rho(p_i, q_j) = 0$. Suppose the positive sample correlation to be such as to be just significant, so that the probability of obtaining this correlation (or one larger) under the null hypothesis is exactly 0.05. Then if the null hypothesis were actually that $\rho(p_i, q_j)$ were strictly negative, say equal to $-\frac{1}{2}$, then the probability of getting a positive sample correlation would be even less, and the observed correlation coefficient more significant statistically. Thus, by having the null hypothesis $\rho(p_i, q_j) = 0$, all i, j rather than $\rho(p_i, q_j) \leq 0$ (or ≥ 0, as the case may be for particular i, j), we obtain conservative hypothesis tests. In other words, the level of significance, under the true null hypothesis is at least as high as it is under the conservative null hypothesis used for testing.

In this fashion, let us establish a simple significance test for the $r(p_i, q_j)$ matrix. If Table 15 specifies that $r(p_i, q_j)$ be positive and the sample coefficient was positive, then that event, the correct prediction of the sign of the sample correlation coefficient occurred under the null hypothesis with probability one-half. But there are sixteen i, j pairs and

for each of them the (marginal) probability of Table 15 predicting the sign correctly, when the null hypothesis is true, is one-half. Now under the assumption that the individual $r(p_i, q_j)$'s are distributed independently, it is easy to test the sign pattern of Table 16.[19]

The probability of predicting at least 14 out of 16 signs correctly under the null hypothesis is simply the probability of obtaining at least 14 successes out of 16 Bernoulli trials (e.g., coin tosses), each with a probability of success of one-half. From tables of the cumulative Binomial distribution, this probability is 0.002. By this test, the sign pattern as predicted by Table 15 is thus strongly statistically significant, and this strong result is conservative if we consider the general null hypothesis to be the appropriate one. The strength of the significance appears to put to rest doubts about the independence

---

[19]This assumption is necessitated by neither the null nor the alternative hypotheses, which allow the $\rho(p_i, p_j)$'s and the $\rho(q_i, q_j)$'s to be nonzero, so the preference and choice variables can be correlated within each set. Furthermore, the assumption does not appear to be supported by the data, as the nonzero p_ip_j and q_iq_j correlations in Table 16 show. Also, even if the $\rho(p_i, p_j)$'s and $\rho(q_i, q_j)$'s were zero, then $r(p_w, q_w)$ and $r(p_w, q_c)$ would be distributed dependently, for example, for the same p_w data are used in the calculation of both.

The assumption is made solely to enable the simple test derivation above, to have some significance test on the whole matrix, without the complexity of the Generalized Likelihood Ratio Test or the Test for Complete Independence. It is unclear how the violation of the assumption affects the significance figures obtained, but it is clear that while the $r(p_i, q_j)$'s are dependently distributed, the dependence is not strong, since none of the correlations in Table 16 is large.

assumption (see note 19). Furthermore, this test has very low power, since it is sensitive only to the sign and not the magnitude of the sample correlation coefficients. In other words, a more powerful test would result an even more strongly significant result.[20]

5.5. Second-Order Preference-Choice Relationships

Using the axioms of Table 14, these results may be extended to the case of several second-order preference-choice relationships. It will also be shown that under certain conditions the scale dependence of the correlation coefficient does not cause as much of a problem as might be expected.

Recall that the standard formula for the product-moment correlation coefficient is

$$(1) \qquad r(p_i, q_j) = \frac{\mathrm{cov}(p_i, q_j)}{(\mathrm{var}(p_i)\mathrm{var}(q_j))^{1/2}}$$

where cov(•,•) and var(•) are sample covariances and variances, if this formula is that for a sample correlation coefficient.

---

[20] The same test may be applied to the 'raw' correlation matrix, that is, the 4 X 15 correlation matrix computed using the component $q_{b,i}$ and $q_{c,i}$ variables. The significance level here is the probability of at least 48 successes in 60 Bernoulli trials, each with success probability one-half. Using the Normal approximation to the binomial distribution a z value of 4.518 is obtained, corresponding to a significance level of less than 10^{-5}. The strong statistical significance of the 4 X 4 matrix is thus repeated here.

Since the denominator of (1) is obviously always positive, the sign of the correlation coefficient is identical to the sign of the covariance, so that Tables 14 and 15 may be interpreted in terms of the signs of cov (p_i, q_j) expressions.

Consider the most general second-order preference-choice correlation. Substituting in (1) find

$$(2)\quad r((p_i-p_j), (q_h-q_k)) = \frac{\text{cov}(p_i, q_h) - \text{cov}(p_i, q_k) - \text{cov}(p_j, q_h) + \text{cov}(p_j, q_k)}{[\text{var}(p_i)+\text{var}(p_j)-2\text{cov}(p_i,p_j)]^{\frac{1}{2}}[\text{var}(q_h)+\text{var}(q_k)-2\text{cov}(q_h,q_k)]^{\frac{1}{2}}}$$

In the case of $r((p_b-p_w), (q_b-q_w))$ substituting in (2) find

$$(3)\quad \text{sgn}[r((p_b-p_w), (q_b-q_w))] =$$

$$\text{sgn}[\text{cov}(p_b, q_b) - \text{cov}(p_b, q_w) - \text{cov}(p_w, q_b) + \text{cov}(p_w, q_w)]$$

where sgn[•] is the operator taking only the <u>sign</u> of the quantity in brackets. But Table 14 specifies the signs of all the covariances on the right side of (3). By axiom I, $\text{cov}(p_b, q_b)$ and $\text{cov}(p_w, q_w)$ are both positive and by axiom IV $\text{cov}(p_b, q_w)$ and $\text{cov}(p_w, q_b)$ are both negative. Thus, by (3) axioms I and IV are <u>sufficient</u> to determine <u>ex hypothesi</u> the sign of $r((p_b-p_w), (q_b-q_w))$, since that quantity is the sum of 4 positive terms.

Consider another second-order correlation. Again substituting in (2) find

(4) $\quad \text{sgn}\,[r((p_b-p_\ell), (q_b-q_\ell))] =$

$$\text{sgn}\,[\text{cov}(p_b, q_b) - \text{cov}(p_b, q_\ell) - \text{cov}(p_\ell, q_b) + \text{cov}(p_\ell, q_\ell)].$$

By axiom I, $\text{cov}(p_b, q_b)$ and $\text{cov}(p_\ell, q_\ell)$ are both positive, axiom III specifies that $\text{cov}(p_b, q_\ell)$ be positive, and axiom IV specifies that $\text{cov}(p_\ell, q_b)$ is negative. However these three axioms are now insufficient to determine the sign of the second-order correlation coefficient, for this sign is determined by the sum of both positive and negative terms. Since our hypotheses do not specify the magnitudes of the covariances, the sign of $r((p_b-p_\ell), (q_b-q_\ell))$ is indeterminate, ex hypothesi.

For the six second-order preference-choice correlations of the form $r((p_i-p_j), (q_i-q_j))$ Table 17a shows the axioms used in calculating the sign of the correlation coefficient, and whether these axioms are sufficient to specify its sign.

While it is certainly convenient to have our first-order axioms specify, in some cases, the signs of the second-order correlations, before this result is used we must show that the sign of these second-order correlations is invariant with scale transformations of the data. In (2), for all the original preference or choice variables consider four positive linear transformations of the form

(5)
$$\begin{aligned} p_i' &= \alpha p_i + \beta \\ p_j' &= \gamma p_j + \delta \\ q_h' &= \lambda q_h + \varepsilon \\ q_k' &= \rho q_k + \sigma \end{aligned} \qquad \alpha, \gamma, \lambda, \rho > 0$$

Table 17. Predicted and Actual Second-Order Preference-Choice Correlations

a. Predicted Qualitative Correlations and Axioms Used in Prediction (see text)

Correlation	I	II	III	IV	Sign Predicted
$r((p_b-p_c),(q_b-q_c))$	X	X			Insufficient
$r((p_b-p_w),(q_b-q_w))$	X			X	+
$r((p_w-p_c),(q_w-q_c))$	X			X	+
$r((p_b-p_\ell),(q_b-q_\ell))$	X		X	X	Insufficient
$r((p_\ell-p_c),(q_\ell-q_c))$	X		X	X	Insufficient
$r((p_w-p_\ell),(q_w-q_\ell))$	X			X	+

b. Actual Correlations, Standardized Data, No Missing Observations (see text)

		No. of Correlation Coefficients:			
Correlation	No. of Observations	Total	With Correctly Predicted Signs	With Correct Signs, Statistically Significant	Average Value of Correlation Coefficients
$r((p_b-p_w),(q_b-q_w))$	329	9	9	0	.049
$r((p_w-p_c),(q_w-q_c))$	329	4	4	0	.115
$r((p_w-p_\ell),(q_w-q_\ell))$	302	1	1	1	.239
Totals		14	14	1	

Then, substituting (5) into (2), find

$$(6) \qquad \text{sgn}\,[r((p'_i-p'_j),(q'_h-q'_k))] =$$

$$\text{sgn}\,[\alpha\lambda\text{cov}(p_i,q_h)-\alpha\rho\text{cov}(p_i,q_k)-\gamma\lambda\text{cov}(p_j,q_h)+\gamma\rho\text{cov}(p_j,q_k)].$$

Now, in situations where the sign of the second-order correlation coefficient is specified by our hypotheses, the numerator of the right side of (2) is the sum of quantities with the same sign. Under these conditions, the sign of the correlation on the transformed variables will also have its sign specified, and have the same sign, for its sign is determined by a differently weighted sum of the same same covariances. Thus, when the axioms of Table 14 (or some of them) are sufficient to determine the sign of a second-order preference-choice correlation, the sign of this correlation is invariant with any positive linear scale transformation. In other words, when its sign is predicted by our axioms, the sign of the second-order preference-choice correlation coefficient is a meaningful quantity, for it is scale invariant.

Notice, however, that the magnitude of the second-order correlation coefficient is influenced by the scales used. Since the test for the statistical significance of a single correlation coefficient is dependent only on sample size, then whether a given sample second-order correlation attains significance is influenced by the scales used.

In Table 17b the actual values are given for the three second-order correlations whose sign is predicted ex hypothesi.

To provide some uniformity in computing these correlations, the q_i and q_j variables have been standardized, that is, transformed linearly to have mean zero and variance unity. In all three cases, the signs were predicted correctly, including all 14 'raw' correlation coefficients. However, as before, the correlation coefficients are small, and only one out of the fourteen attains individual statistical significance, as has been defined.

Applying the binomial significance test of the preceding section, we calculate the probability of 14 successes in 14 Bernoulli trials, each with a success probability of one-half. From tables of the binomial distribution, this probability is less than 0.0005. This strong statistical significance is obtained using correlation coefficients computed from the component $q_{b,i}$ and $q_{c,i}$ variables.[21] Notice that the binomial significance test is especially appropriate for the second order preference-choice relationship. Here the scales used affect the magnitude, but not the sign of the (population)correlation coefficient, while the binomial test is sensitive to the sign, but not the magnitude of the (sample) correlation coefficient. Thus test significance level is scale-invariant.

---

[21]If the correlation coefficients in each row of Table 17 are averaged, an operation paralleling that creating Table 16, then there are 3 correlation coefficients, all of whose signs were correctly predicted, and with one statistically significant. This is a 'perfect score' but since there are but 3 coefficients, the binomial test probability associated with this result is 0.125, failing statistical significance.

In conclusion, the second-order preference-choice correlations support our axioms, just as did the first-order correlations. Notice in Table 17a, however, that just axioms I and IV are being tested.

6. SUMMARY OF RESULTS, CONCLUSIONS, AND SUGGESTIONS FOR FURTHER RESEARCH

In spite of the many difficulties plaguing this research, it does appear that useful results have been obtained. The difficulties include those of measuring and even defining preference and choice variables, the unusual areal composition of the sample, and the two-variables-at-a-time form of analysis which excludes many important influences and contributes to the small correlation coefficients.

'Useful results' mean more than the attainment of statistical significance. More important is the meaningful qualitative pattern of correlations obtained among the preference variables, the choice variables and (especially) between the preference and choice sets.

6.1. Results with Preference and Choice Variables

In general, the preference variables (in both Sections 2 and 5) are either negatively correlated or unassociated with each other. An exception is the positive association of preferences for natural environmental beauty (B and p_b) with those for the nonnatural environment in general, and low density surroundings in particular (M and p_c). This pattern of intercorrelations suggests that accessibility, house and lot characteristics, and natural and nonnatural environmental characteristics (considered as one) are 'substitute residential commodities' in the sense that

a preference for one of them reduces (or at least does not increase) the individual's preference for another. Since these preference categories were selected on a prior basis, the pattern of intercorrelations provides some statistical confirmation for them, that is, positively associated preference categories would have an element of redundancy, while the present categories do not.[22]

Another result (see Table 5) is that in the Brandywine, Neshaminy, Northern Chester County, and Main Line areas, the ordering of the four preference categories (starting with the most frequently mentioned one) is characteristics of house and lot (H), accessibility (A) and man-made environmental characteristics (M), (A and M considered as one), followed by natural environmental beauty (B). However this preference ordering is to be interpreted with great caution, as discussed in subsection 2.3. There do not seem to be very large interareal preference differences.

The residential choice measures (see Section 4 and subsection 5.2) also appear to have good properties and lead to useful results. The various measures of natural environmental quality (the $q_{b,i}$ variables) are generally positively associated with one another, and have the proper direction of association with the preference variables, intuitively suggesting that they are all measuring aspects of the same

---

[22]This indicates, in passing, one advantage of a prior categorization. Many of the common statistical approaches would automatically result in uncorrelated variables (e.g., principal component and factor analysis), denying us the very useful interpretation of negatively associated variables as representing substitute commodities.

thing, and that this 'thing' may well be the quality of the natural environment. Similar results are obtained with variables defined on both the half and one mile circle, suggesting that the problem of the proper areal unit may perhaps not be crucially important in practice, at least at this scale of analysis, unless one hopes for much larger correlations than those obtained here. However, while the specific $q_{b,i}$ variables are all positively associated, this intercorrelation is not strong, in fact only as strong as the relationship with the $q_{c,i}$ avoidance of density and congestion measures. This is one indication, among many, that the present research is not sensitive enough to separate natural environmental beauty per se from the absence of homes and other signs of human activity.

Furthermore, the results obtained with the choice variables suggests that, at least at this scale of analysis where we are not interested in the varieties of natural environmental quality, much can be accomplished without taking natural environmental perception into account.

Especially gratifying is the fact that, within our sample, the intercorrelations among the choice variables take the same qualitative form as would be expected throughout the metropolitan area as a whole (see Table 16 and subsection 5.2). For example, a residence's accessibility is negatively associated with its lot size or, the larger homes are further out. This pattern of intercorrelations not only helps support inferences from our sample to the metropolitan area as a whole, but

suggests that, for example in choosing a residence respondents have to trade off between accessibility and lot size. This further substantiates our notion of the residential attributes as substitute commodities, and preferences as tradeoffs among them.

Of course, the most important set of correlations is that between the preference and choice variables. Here, given four axioms (Table 14) as to individual preference-choice correlations, we can predict the signs of all the first-order correlations between absolute levels of preference and choice variables (Table 15) and some of the second-order correlations between the relative levels of pairs of preference and choice variables. These predictions are substantiated by the data. While the magnitudes of the correlation coefficients are quite small, the matrices as wholes are highly significant statistically. Among the four axioms are the manifestation of preferences in choice (I), and the substitutability of most different residential attributes (IV); in fact these two axioms alone result in the 3 second-order preference-choice relationships of Table 17.

Now mathematical formulations with several variables are quite common in mathematical social science. Rarely in those cases can we predict by theory the magnitudes of correlation or regression coefficients, or partial derivatives, but we can frequently predict some or all of the signs. In those cases a statistical analysis like that of Section 5 should be useful, where the hypothesis test refers to the qualitative form of

the entire matrix, not just individual elements. Hopefully, a more powerful test than the binomial test on signs can be developed, that, like the binomial test, is 'one-tailed' for it would be sensitive to the signs of the predicted and actual matrix elements.

One result of subsections 5.3 and 5.5 would seem to have relevance to preference studies in general. While second-order preference-choice correlations are often more intuitively appealing than first-order correlations, the former are sensitive to the scales used which here, as in most studies with nonstandardized commodities, are inherently arbitrary. The mathematical result of subsection 5.5 is that, for certain combinations of the signs of first-order preference-choice correlations, the sign of the second-order correlation is scale-invariant, that is invariant up to positive linear transformations. In these situations not only is the sign of the second-order correlation a meaningful entity, since it is scale-invariant, but that sign may be deduced from the signs of first-order relationships.

It would be useful if these results could be extended to nonlinear preference-choice relationships and general (that is, monotonic) scale transformations.

6.2. Socio-Economic and Demographic Influences on Preferences

Compared to the preference-choice relationships, the attempts to associate preferences with socio-economic and demographic influences yielded overwhelmingly negative results.

The problems cited in subsection 3.5, of imperfect measures, two-variables-at-a-time analysis, and statistical tests with low power all occurred in Section 5, yet positive results were obtained there. One conclusion, offered hesitantly, is that the 'standard' socio-economic and demographic explanatory variables are less useful in explaining residential preferences than is common. Similarly, the location of the respondent's childhood residence, usefully applied by Sonnenfeld [24], yielded no positive results here. It may be possible that a person's preferences (as we have defined them) are better predicted by the characteristics of his present residence, than his socio-economic and demographic attributes. However, to establish that conclusion, we would have to use the same mathematical form, say a regression equation, to compare the relative explanatory power of the choice variables and the socio-economic and demographic variables. Perhaps, though, if inner city residents were sampled, so there was a greater range of socio-economic variation, then some preference relationship would be demonstrated.

6.3. Demand-Oriented Theories of Urban Form

The existence of a definite preference-choice relationship adds some empirical credence to demand-oriented theories of urban form, as discussed in subsection 0.3, especially since the relationships are simple enough to be deduced from the four axioms of Table 14. As was mentioned, the small magnitudes of the correlations are better interpreted as due to the imperfections of our variables, than due to the possibly low strength of the relationships assumed by these urban form theories.

The simple substitute relationships among the various 'residential commodities' may, however, help in the construction of such theories. Simpler and more easily manipulated utility functions, leading possibly to stronger results, might be constructed using the assumption of substitutability. Frequently, (see, e.g., Alonso [1]) these theories consider only the residential attributes of space and accessibility. Perhaps our results make possible the construction of a manipulable utility function also defined on the residence's natural and non-natural environmental characteristics, though this task still appears quite difficult.

6.4. Evaluating a Cluster Development Pattern

An initial motivation for this research (see subsection 0.1) was to assess persons' desire for a cluster development residential bundle, as opposed to the more common suburban sprawl combination. Both Tables 6 and 16 indicate that the preference for house and lot characteristics (or simply a large lot) is significantly negatively associated with preferences for natural environmental beauty, so there presumably are persons willing to sacrifice lot size to achieve the beautiful natural environment of the Brandywine cluster plan. However, both tables show the significantly positive association between natural environmental preferences and non-natural environmental preferences and (for Table 16) preferences for the avoidance of high density surroundings. Thus, persons who want to live in a high quality natural environment tend also to want low density surroundings. (See also axiom II, Table 14.) Since

cluster development would raise the nearby population densities over those of the spread pattern (because of smaller lot sizes) it is hard to use this part of our results to evaluate people's reaction to the cluster pattern.

Part of this difficulty may come from the fact that our measures are probably insensitive to differences between 'nearby' and 'overall' population densities. Under a cluster pattern, overall density is likely to be similar to that under the spread pattern. Indeed, an early working hypothesis, not substantiated by the data, was that natural environmental quality and low density are separate preference items, just like the others. The hypothesis was suggested by interviews with rural residents who carefully distinguished between the two: some were interested in only a pretty environment, others in just getting away from other people. Most suburbanites, on the other hand, are unlikely to have had experience with areas where the two variables vary separately, and are likely to perceive of natural environmental quality as identical to low population density.

To make some a sessment of new residential bundles, though, we must reason quantitatively, using the magnitude of various preference-preference and preference-choice relationships, as well as their signs. Furthermore, we must have some indication of the numbers of persons holding the various preferences. Thus, if 80% of the persons had a strong preference for a large lot, and but 10% strongly preferred a pretty natural environment, then even a very high negative correlation between the two preferences might not indicate much demand for a

cluster development pattern.

Another interesting result (see subsection 5.4 and axiom III, Table 14) is that natural environmental preferences tend to manifest themselves by increased lot sizes. One explanation for this is the uncertainty people have about whether the quality of their surroundings will endure. This reasoning suggests that land use controls that are effective and known to be effective in environmental preservation would assure the continuance of a given level of environmental quality and would encourage people to buy smaller lots which would in turn reduce the quantity of rural land converted to residential uses, and the accompanying environmental damage.

6.5. Suggestions for Further Research

Clearly, it would be desirable to work with a sample that is representative of the entire metropolitan area, and with more refined preference and choice measures. It would also be interesting to have family income included in the preference-choice relationships, especially the first-order ones. Either the correlations could be computed separately for different income classes, or income could be included as a second explanatory variable. First of all, we might see if richer persons are better able to express their preferences in the housing and space markets. Secondly, whatever their preferences, it is clear that wealthy persons can afford more of all residential attributes than poor persons, so the level of a choice variable should be influenced by income as well as preferences.

Similarly, we might experiment with including respondents' information about housing opportunities as measured by, say, the number of homes they looked at before choosing the present one, and whether they had lived before in the same neighborhood as their present home.

Another possibility is to include the length of stay in the present residence in the preference-choice relationship. One would expect the preference-choice relationship to be strongest for new and recent movers, as has been mentioned (subsection 5.3). At the same time, calculating preference-choice relationships separately for those asked about the choice of the present home, and about a hypothetical home (see subsection 2.1) might evaluate one possible confounding influence on the preference-choice relationship. Suppose there were no relationship between 'true' preferences and choice, but that our 'preference' measures were strongly influenced for recent movers by the characteristics of the present residence as well as by true preferences. This could happen since the question asks respondents about the considerations that were important in the choice of the present residence. In this situation our 'preference' measures could be spuriously associated with residential choice for recent movers, but not for those asked about the choice of a hypothetical home.

Finally, non-empirical future research might concentrate on deducing the mathematical form of first- or second-order preference-choice relationships from the Theory of Consumer Behavior. Some initial encouraging results have been obtained

by assuming explicit functional forms for the utility function. Linear choice-preference relationships have been obtained with these utility functions, with income appearing in the first-order relationship but not the second-order one. One important task in this analysis is to interpret our notion of preferences in terms of the parameters in explicit utility functions or, if results may be obtained using abstract utility functions, in terms of the partial derivatives.

Theoretical deduction of preference-choice relationships not only justifies empirical preference-choice analysis, but may perhaps indicate how preference data like those here may be used to estimate utility and demand functions. Furthermore, this work might help add natural environmental and density considerations to the utility functions in the demand-oriented urban form theories.

REFERENCES

[1] Alonso, W. Location and Land Use. Cambridge: Harvard University Press, 1965.

[2] Coughlin, R. E. and K. A. Goldstein. "The Extent of Agreement Among Observers on Environmental Attractiveness," Discussion Paper Number 37, Regional Science Research Institute, Philadelphia, 1970.

[3] Duncan, O. D., R. P. Cuzzort, and B. Duncan. Statistical Geography: Problems in Analyzing Areal Data. New York: The Free Press, 1961.

[4] Gans, H. J. "Urbanism and Suburbanism as Ways of Life: A Re-Evaluation of Definitions," in A. M. Rose (ed.) Human Behavior and Social Processes. Boston: Houghton Mifflin, 1962, pp. 625-648.

[5] _____. The Levittowners. New York: Pantheon Books, 1967.

[6] Harris, B., J. Nathanson and L. Rosenburg. Research on an Equilibrium Model of Metropolitan Housing and Locational Choice, Interim Report. Institute for Environmental Studies, University of Pennsylvania, 1966.

[7] Henderson, J. M. and R. E. Quandt. Microeconomic Theory: A Mathematical Approach. New York: McGraw-Hill, 1958.

[8] Herbert, J. D. and B. H. Stevens. "A Model for the Distribution of Residential Activity in Urban Areas," Journal of Regional Science, 2 (1960), 21-36.

[9] Institute for Environmental Studies, University of Pennsylvania; Regional Science Research Institute; and U.S. Geological Survey, The Plan and Program for the Brandywine. Philadelphia: Institute for Environmental Studies, University of Pennsylvania, 1968.

[10] Keeping, E. S. Introduction to Statistical Inference. Princeton: Van Nostrand, 1962.

[11] Kendall, M. G. Rank Correlation Methods. London: Griffin, 1948.

[12] Lansing, J. B. and E. Mueller. Residential Location and Urban Mobility. Ann Arbor, Michigan: Institute for Social Research, 1964.

[13] Luce, R. D. and H. Raiffa. Games and Decisions. New York: Wiley, 1957.

[14] McClelland, D. The Achieving Society. Princeton: Van Nostrand, 1961.

[15] Menchik, M. "Brandywine Residents and Their Attitudes towards the Natural Environment," in [9], Chapter II-G and Appendix pp. 21-48.

[16] Morrison, D. F. Multivariate Statistical Methods. New York: McGraw Hill, 1967.

[17] Muth, R. F. Cities and Housing. Chicago: University of Chicago Press, 1969.

[18] Pearson, E. S. and H. O. Hartley (eds.). Biometrika Tables for Statisticians, Vol. 1, 3rd ed. Cambridge: Cambridge University, 1966.

[19] Peterson, G. L. "A Model of Preference: Quantitative Analysis of the Perception of the Visual Appearance of Residential Neighborhoods," Journal of Regional Science, 7 (1967), 19-32.

[20] Quandt, R. E. and W. J. Baumol. "The Demand for Abstract Transport Modes: Theory and Measurement," Journal of Regional Science, 6 (1966), 13-26.

[21] Rabinowitz, C. B. and R. E. Coughlin. "Analysis of Landscape Characteristics Relevant to Preference," Discussion Paper Number 38, Regional Science Research Institute, Philadelphia, 1970.

[22] Rossi, P. H. Why Families Move. New York: The Free Press, 1955.

[23] Siegel, S. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill, 1956.

[24] Sonnenfeld, J. "Variable Values in Space and Landscape: An Inquiry into the Nature of Environmental Necessity," in R. W. Kates and J. F. Wohlwill (eds.), "Man's Response to the Physical Environment," The Journal of Social Issues, 22 (1966), 71-82.

[25] Waugh, E. "An Englishman's Home," in Work Suspended and Other Stories. Penguin: Harmondsworth, England, 1951, pp. 42-61.

[26] Colby, B. N. and M. Menchik. "A Study of Thematic Apperception Tests with the General Inquirer System," El Palacio, (Winter, 1964), 29-36.

RECENT DISCUSSION PAPERS OF THE
REGIONAL SCIENCE RESEARCH INSTITUTE

No. 29 IMPACT OF VIET NAM WAR EXPENDITURES ON THE PHILADELPHIA ECONOMY, by Walter Isard and Thomas W. Langford, Jr., May 1969. $3.00.

No. 30 GROWTH POLES: AN INVESTIGATION OF THEIR POTENTIAL AS A TOOL FOR REGIONAL ECONOMIC DEVELOPMENT, by Vida Nichols, May 1969. $2.50.

No. 31 TRENDS IN INDUSTRIAL LOCATION AND THEIR IMPACT ON REGIONAL ECONOMIC DEVELOPMENT, by Benjamin H. Stevens, Robert C. Douglas and Carolyn B. Neighbor, May 1969. $2.00.

No. 32 ON THE LÖSCHIAN SPATIAL DEMAND CURVE, by Colin A. Gannon, April 1969. $2.00.

No. 33 INDIVIDUAL EQUILIBRIUM-SEEKING BEHAVIOR: A DEFINITION AND A TEST, by Mark D. Menchik, June 1969. $3.50.

No. 35 ECOLOGICAL STATISTICAL STUDIES ON ENVIRONMENTAL POLLUTION AND CHRONIC DISEASE IN METROPOLITAN AREAS OF THE U.S., by Richard J. Hickey, David E. Boyce, Evelyn Harner, and Richard C. Clelland, February 1970. $3.50.

No. 36 CRITERIA FOR MEASUREMENT OF STREAM CHANNELS AS AN INDICATOR OF PEAK FLOW HISTORY, by Thomas R. Hammer, February 1970. $2.50.

No. 37 THE EXTENT OF AGREEMENT AMONG OBSERVERS ON ENVIRONMENTAL ATTRACTIVENESS, by Robert E. Coughlin and Karen A. Goldstein, February 1970. $2.50.

No. 38 ANALYSIS OF LANDSCAPE CHARACTERISTICS RELEVANT TO PREFERENCE, by Carla B. Rabinowitz and Robert E. Coughlin, March 1970. $3.00.

No. 39 AN ECONOMETRIC FORECASTING MODEL FOR THE PHILADELPHIA REGION, by Norman J. Glickman, May 1970. $2.50.

No. 40 SOLVING URBAN LOCATION PROBLEMS: HUMAN INTUITION VERSUS THE COMPUTER, by Jerry B. Schneider, June 1970. $1.50.

RECENT DISCUSSION PAPERS (Cont.)

No. 41 GOAL ATTAINMENT LEVELS IN 101 METROPOLITAN AREAS, by Robert E. Coughlin, December 1970. $3.00.

No. 42 AIR QUALITY AND ITS RELATIONSHIP TO ECONOMIC, METEOROLOGICAL, AND OTHER STRUCTURAL CHARACTERISTICS OF URBAN AREAS IN THE UNITED STATES, by John R. Welson and Benjamin H. Stevens, December 1970. $2.50.

No. 43 SOME EXPERIMENTS IN QUANTITATIVE MEASUREMENT OF LANDSCAPE QUALITY, by Carla B. Rabinowitz and Robert E. Coughlin, March 1971. $2.00.

No. 44 DUALITY THEORIES IN LINEAR, QUADRATIC AND CONVEX PROGRAMMING: A SURVEY, by Nazir G. Dossani, April 1971. $2.00.

No. 45 LAND VALUES AND ENVIRONMENTAL CHARACTERISTICS IN THE URBAN-RURAL FRINGE, Robert E. Coughlin and James Fritz, May 1971. $2.00.

No. 46 RESIDENTIAL ENVIRONMENTAL PREFERENCES AND CHOICE: SOME PRELIMINARY EMPIRICAL RESULTS RELEVANT TO URBAN FORM, by Mark D. Menchik, March 1971. $3.00.

Regional Science Research Institute Discussion Papers are available on a prepaid subscription basis ($18.00 per year for an average of eight papers). A limited number of copies of current issues is also available for single purchases.